COARSE-TO-FINE MULTIPLE TESTING STRATEGIES.

Under the direction of Dr Donald Geman and Dr Laurent Younes.

by

Kamel Lahouel

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

February, 2018

Abstract

We consider a multiple testing scenario encountered in the biological sciences and elsewhere: there are a great many null hypotheses about the distribution of a high-dimensional random variable but only a very small fraction are false (or "active"); moreover, controlling the false positives rate through FWER or FDR is imperative. Not surprisingly, the usual methods applied to control the two former criteria are often too conservative and lead to a small number of true detections. Clearly, some additional assumptions or domain-specific knowledge are then necessary to improve power. Motivated by applications in genomics, particularly genome-wide association studies, we suppose the set indexing the hypotheses has a natural hierarchical structure, the simplest case being a partition into "cells." In principle, it should then be possible to gain power if the active hypotheses tend to cluster within cells. We explore different coarse-to-fine, two-level multiple testing strategies, which control the FWER or the FDR and are designed to gain power relative to usual single level methods, in so far as clustering allows it. Simulations confirm a sharp improvement for in data models we consider.

ABSTRACT

Acknowledgments

I owe a lot of gratitude to Professors Donald Geman and Laurent Younes for their great help and encouragement throughout this research work. Without their valuable advice, insights, challenging interaction and continuous support, this thesis would not have taken shape. I have been lucky and privileged to benefit from their rich research and teaching experience.

Dedication

I dedicate this thesis to my wife for her moral support and full understanding. My thanks also go to my parents for their relentless encouragement and push for perseverance and for my brother and sister for being on my side when I need them. I also dedicate this thesis to my grandmother whose strong moral support and affection have been a source of strength for me. Finally, I have a special thought for my friends and extended family for stimulating discussions and their joyful company.

Contents

List of Tables

List of Figures

Chapter 1

Introduction and background

1.1 A high-level introduction

Multiple hypotheses testing problems, sometimes called multiple comparisons problems, are encountered when one needs to test a number of statistical hypotheses, and accepts or rejects each one of them. The final output of this procedure has the form of a subset, representing the subset of rejected null hypotheses. One can view a multiple testing procedure as a strategy allowing to answer multiple yes/no questions. Common sense suggests that the more questions we ask and try to answer, the larger will be the number of sources of errors. Moreover, if these questions are of the yes/no type and we want to avoid making one type of incorrect answers, say for example answering yes while the correct answer is negative , one will tend to systematically answer negatively to almost all of the questions.

In fact, imagine a game with the following setting:

- ⊙ A player is put in front of 60 boxes that can be labeled with either 0 or 1.
- ⊙ He is told that each of the boxes is either empty or contains a certain amount of money.
- ⊙ If he opens any empty box or misses a box containing money, he needs to pay a certain amount; otherwise he wins the amount of money in the boxes if and only if he opens all the boxes containing money and only those boxes.
- ⊙ The player is given the information that if a box is empty, its label has been generated such that the probability of labeling it with 0 is 0.001. Moreover, all boxes with money inside are labeled with the number 0.
- ⊙ Once the labels revealed, the player can decide whether or not to open each box.
- ⊙ Lastly, he knows that in case he looses the game, he will have to pay 19000 dollars. If he wins, the total amount of money he will find in the boxes is 1000 dollars.

Considering the potential gains and losses, the player decides that he needs to control the probability of loosing at a level less than 0.05; otherwise the game is not worth playing. He obviously decides that in case he opens boxes having a certain label, this label will necessarily be 0. He then computes the probability of opening any

empty box which is is guaranteed to be controlled at approximately 0.06. Therefore, he comes to the conclusion to leave the game before it is too late! Generally, as long as the number of boxes is greater than 50, the player shouldn't open any box with his requirement that the probability of opening any empty box should be less than 0.05. Moreover he cannot decide to consider a smaller subset of the boxes prior to seeing the labels, as he needs to make sure to open all the boxes containing money.

A similar situation is frequently encountered in multiple hypothesis testing. Usually, one needs to control the false positives in a specific way. The larger is the number of hypotheses, the stronger is the evidence required to reject the null hypotheses. The strength of this evidence, which corresponds to the label of the box in the previous example, depends on the number of observations that we have and, very frequently, we face a situation where the number of tested hypotheses is large enough compared to the number of observations, to make the number of detections considerably lower than the actual number of true positives. To address this issue, the work proposed in this thesis will consist of procedures assuming that the false null hypotheses tend to cluster together. As a result these procedures will be less conservative provided our assumption is correct while controlling the false positives at the desired level. A legitimate question to ask is whether we could still control the false positives if the clustering assumption made is incorrect. The answer to this question is positive, although the power of detecting true positives in that situation will be lower.

To see the big picture, let us go back to our game. Assume now that the boxes are

placed in groups of four, so that we have 15 groups of 4 boxes. The player is given the information that all boxes in a same group are of the same type: either they are all empty or they all contain a certain amount of money. Being a smart player, he notices that all he needs to do is to choose a predetermined representative box for each of the 15 groups so that he looks at 15 boxes in total. His strategy is then the following: for each of these boxes, if the label of the box is 1, then don't open all the boxes in the same group. If the label of the box is 0, then open all the boxes in the same group. This strategy guarantees a probability of winning the game of at least 98.5% and the game becomes worth playing. In this situation, the clustering information is a game changer. In reality, the situation is not as extreme in the multiple testing context. The first reason is obviously that it is generally impossible to guarantee detecting all the true positives. The second reason is that it is not possible to be sure about such kind of clustering assumptions. To maintain boxes game analogy, imagine now that the player will need to pay money only if he opens an empty box. He decides to play the game only if he is sure to control the probability of having to pay a certain amount at a level less than 5%. The clustering information is no longer certain, i.e more of a rumor that he heard from the organizers of the game but it is not guaranteed that it is a correct information. The player notices the following: What if I only decide to open the 15 representatives from of each group? If the clustering rumor is correct, then I'll open exactly 25% of the boxes containing money which is a decent amount. If it is incorrect, I might not win as much money, but in any case I control the probability

of having to pay money. Therefore I'm going to play the game anyway!

The family of multiple testing procedures that we propose in this thesis will present analogous behavior: The procedures will be designed and adapted for situations where the clustering assumption is true, but they will control the false positives regardless. Even if the last strategy described in our imaginary game can be seen as a coarse-to-fine strategy, it is radical in the sense that it throws away 75% of the boxes without even looking at them. Our procedures will more closely resemble the following pattern: for each group, first look at all the labels of all the boxes as a total information and then decide either to keep all the boxes or throw them away. After that first filter, look individually at the label of the remaining boxes and decide whether or not to open them. However, controlling false positives in this situation is mathematically more challenging than the strategy described in the boxes game. The bulk of this work is to prove that the proposed strategies control the false positives in a sense that will be defined. For this, we first need to formally pose the problem.

1.2 A short introduction to hypothesis testing

A statistical model is a triple $(\mathcal{U}, \mathcal{F}, \mathbb{P}_{(\mathbb{P}\in\mathcal{P})})$ where $\mathcal{U}$ is a set, usually of the form $\mathbb{R}^{n\times m}$, $\mathcal{F}$ is a σ-algebra on $\mathcal{U}$ and $\mathbb{P}_{(\mathbb{P}\in\mathcal{P})}$ is a family of probability measures defined on the measurable space $(\mathcal{U}, \mathcal{F})$ indexed by a set of probability measures $\mathcal{P}$. A

statistical hypothesis testing can be viewed as follows. One observes a random vector $\mathbf{U}$ belonging to the set $\mathcal{U}$ that has been generated by an unknown probability law $\mathbb{P} \in \mathcal{P}$. The goal is to use that observation in order to determine whether $\mathbb{P}$ belongs to a certain subset $\mathcal{P}_0 \subset \mathcal{P}$ or to its complement in $\mathcal{P}$ denoted by $\mathcal{P}_1$. The first hypothesis is called null hypothesis, usually denoted by H_0 and the second the alternative, usually denoted by H_1. Formally, one needs to design a measurable binary function $\phi : \mathcal{U} \to \{0, 1\}$ where 0 corresponds to deciding that $\mathbb{P}$ is in $\mathcal{P}_0$ and 1 to deciding that it is in $\mathcal{P}_1$. Clearly, our decision can be incorrect in two ways: either the unknown probability $\mathbb{P}$ came from the set $\mathcal{P}_0$ and $\phi(\mathbf{U}) = 1$, or the true probability belongs to $\mathcal{P}_1$ and $\phi(\mathbf{U}) = 0$. The probability of making the first kind of incorrect decision is called the type 1 error, and the second one the type 2 error. It is easy to see that if the supports of $\mathcal{P}_1$ and $\mathcal{P}_0$ are not disjoint, it is not possible to design a test with any arbitrarily fixed type 1 and type 2 errors. Therefore, one needs to prioritize one type of errors over the second one, in the sense that we fix an arbitrary level of error for a chosen type that cannot be exceeded by the test. The convention is to fix the level for the type one error. One needs, for a fixed $\alpha \in [0, 1]$:

$$\mathbb{P}\left(\phi(\mathbf{U}) = 1\right) \leq \alpha, \forall\ \mathbb{P} \in \mathcal{P}_0.$$

Therefore, the null hypothesis should be chosen such that one is conservative towards the null and skeptical towards the alternative. We can then compare dif-

ferent tests controlling the type 1 error at a fixed level by comparing their respective type 2 errors [1]. A famous example is when we observe two different i.i.d populations $X_1, X_2, ...X_n$ and $Y_1, Y_2, ..., Y_m$ coming from two normal distributions $\mathcal{N}(\mu_x, \sigma^2)$ and $\mathcal{N}(\mu_y, \sigma^2)$, and want to decide if μ_1 is different from μ_2. In this setting, $\mathbf{U} = (X_1, X_2, ..., X_n, Y_1, Y_2, ...Y_m)$ and $\mathcal{U} = \mathbb{R}^{n+m}$. Here $\mathcal{P} = \{p_{\mu_1,\sigma}^{\otimes n} \otimes p_{\mu_2,\sigma}^{\otimes m} : \mu_1, \mu_2 \in \mathbb{R}, \sigma > 0\}$ and $\mathcal{P}_0$ is the set of probabilities such that $\mu_1 = \mu_2$. An appropriate test for this setting is the well known t-test:

$$T(\mathbf{U}) = \frac{\bar{X} - \bar{Y}}{s\sqrt{(\frac{1}{n} + \frac{1}{m})}},$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ and $s = \sqrt{\left(\frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}\right)}$, with $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ and $s_y^2 = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \bar{Y})^2$. One can easily show that for every $\mathbb{P} \in \mathcal{P}_0$, $T(\mathbf{U})$ follows a Student's t distribution with $n + m - 2$ degrees of freedom. Therefore, a reasonable choice to control the type 1 error at a particular level, consists of defining $\phi(\mathbf{U})$ as $\mathbf{1}_{T(\mathbf{U})\in\Gamma}$, where Γ is a subset of $\mathbb{R}$ such that $\mathbb{P}(Z \in \Gamma) \leq \alpha$ for any random variable Z having a Student distribution with $n + m - 2$ degrees of freedom. Starting from this particular example, a legitimate question to ask is the following: Imagine that instead of 2 populations, we had N populations $(X_{1,1}, X_{1,2}..., X_{1,n_1})$, $(X_{2,1}, X_{2,2}..., X_{2,n_2})$,..., $(X_{N,1}, X_{N,2}..., X_{N,n_N})$, and we would like to compare the pop-

[1] In practice the alternative hypothesis is composite. In some cases, one can fix a subfamily of alternative hypotheses and consider the worst case or consider a sequence of alternatives converging to the worst case probability among the alternative hypotheses

ulations pairwise. Clearly, we have multiple questions and multiple decisions to make. The number of ways in which one can make an incorrect decision is also certainly larger. Would it make sense to talk about type 1 and type 2 errors? The multiple testing framework formalizes this kind of settings.

1.3 The multiple testing framework and background.

Here also, we consider a statistical model $(\mathcal{U}, \mathcal{F}, \mathbb{P}_{(\mathbb{P}\in\mathcal{P})})$ and we observe a random vector $\mathbf{U}$ coming from an unknown probability distribution $\mathbb{P}$ belonging to the family $\mathcal{P}$. The difference is that we are not asking a single question anymore about the probability $\mathbb{P}$. Instead, we consider a set V indexing $|V|$ (cardinality of the set) number of questions where for each $v \in V$, we are asking whether or not $\mathbb{P}$ belongs to a subset $\mathcal{P}_v$ of $\mathcal{P}$. To these questions corresponds a family of null hypotheses $(H_0(v), v \in V)$. Evidently, for each question, we need to design a test appropriate for that particular question v, namely $\phi_v(\mathbf{U}) : \mathcal{U} \to \{0, 1\}$. $\phi_v(\mathbf{U})$ will generally have the form $\mathbf{1}_{T_v(\mathbf{U})\in\Gamma_v}$ for some family of statistics $(T_v, v \in V)$ and rejection regions $(\Gamma_v, v \in V)$. To go back to the example of comparing N populations pairwise, the set V would be the set of subsets $\{i, j\}$ with $i \neq j$ of subsets of size 2 of $\{1, 2, 3, ...N\}$, so that $|V| = \frac{N(N+1)}{2}$. The random vector $\mathbf{U}$ will be $(X_{1,1}, X_{1,2}..., X_{1,n_1})$, $(X_{2,1}, X_{2,2}..., X_{2,n_2})$,..., $(X_{N,1}, X_{N,2}..., X_{N,n_N})$ so that $\mathcal{U}$ is $\mathbb{R}^{\sum_{i=1}^{N} n_i}$. The family $\mathcal{P}$ is $\mathcal{P} = \{\otimes_{i=1}^{N} p_{\mu_i,\sigma}^{\otimes n_i} : \mu_i \in$

$\mathbb{R}\ \forall i \in \{1, ..., N\}, \sigma > 0\}$. The questions we are asking are therefore for each $v = \{i, j\} \in V$, whether or not $\mu_i = \mu_j$, and $\mathcal{P}_v$ are the $\mathbb{P} \in \mathcal{P}$ such that $\mu_i = \mu_j$. Finally, one choice of the family $(T_v, v \in V)$ could be [2] the following for $v = \{i, j\}$:

$$T_v(\mathbf{U}) = \frac{\bar{X}_{i,\cdot} - \bar{X}_{j,\cdot}}{s_{i,j}\sqrt{(\frac{1}{n} + \frac{1}{m})}},$$

where $\bar{X}_{i,\cdot} = \frac{1}{n_i}\sum_{k=1}^{n_i} X_{i,k}$, $\bar{X}_{j,\cdot} = \frac{1}{n_j}\sum_{k=1}^{n_j} X_{j,k}$ and $s = \sqrt{\left(\frac{(n_i-1)s^2_{i,\cdot}+(n_j-1)s^2_{j,\cdot}}{n_i+n_j-2}\right)}$, with $s^2_{i,\cdot} = \frac{1}{n_i-1}\sum_{k=1}^{n}(X_{i,k} - \bar{X}_{i,\cdot})^2$ and $s^2_{j,\cdot} = \frac{1}{n_j-1}\sum_{k=1}^{n}(X_{j,k} - \bar{X}_{j,\cdot})^2$. For every v, the statistic $T_v(\mathbf{U})$ follows a t-distribution with $n_i + n_j - 2$ degrees of freedom whenever $\mathbb{P}$ is in $\mathcal{P}_v$. It is therefore possible to control $\mathbb{P}\left(T_v(\mathbf{U}) \in \Gamma_v\right)$ for every $\mathbb{P}$ that is in $\mathcal{P}_v$. However, the number of ways to make an incorrect decision in this setting is $(2^{|V|} - 1)2^{|V|}$. Hence it makes no sense to talk about such a thing as type 1 and type 2 errors. How should we therefore control the probability of making incorrect decisions.

A natural way to look at this problem is the following: Define the set $A \subset V$ of the indices $v \in V$ such that $\mathbb{P} \notin \mathcal{P}_v$, or in other words the set of indices over which the null hypotheses are false. In this thesis, we will refer to A as "the active set". On the other hand, define $\hat{A}(\mathbf{U})$ as the estimator of A based on the random vector $\mathbf{U}$:

$$\hat{A}(\mathbf{U}) = \{v \in V : \phi_v(\mathbf{U}) = 1\}.$$

[2] This is definitely not the best choice of statistics under this setting. Here, the purpose is to introduce the multiple testing framework in the simplest way. We refer to[2] for the appropriate statistics to use.

Here, we will refer to $\hat{A}(\mathbf{U})$ as detections or discoveries. With these definitions, we would like to control in some way the size of the intersection between the detection set and the non-active set, i.e the complement of of the active set in V. This size can be expressed as $|\hat{A}(\mathbf{U}) \cap A^c|$. Naturally, the goal of a multiple testing procedure is to control the former quantity and maximize at the same time the number $|A \cap \hat{A}(\mathbf{U})|$ of detected true positives. There are two widely used criteria for controlling false positives :

FWER: Assume that $\mathbf{U}$ is defined on the probability space $(\Omega, \mathbb{P})$. The family-wise error rate (FWER) is

$$\mathrm{FWER}(\hat{A}) = \mathbb{P}\left(\hat{A}(\mathbf{U}) \cap A^c \neq \emptyset\right),$$

which is the probability of making at least one false discovery. This is usually controlled using Bonferroni bounds and their refinements,[3–6] or using resampling methods or random permutations. To set the stage, suppose we are given a family $\{T_v = T_v(\mathbf{U}), v \in V\}$ of test statistics and can assume that deviations from the null are captured by small values of $T_v(\mathbf{U})$ (e.g., p-values). Assume that individual rejection regions are of the form $\{u \in \mathcal{U} : T_v(u) \leq \theta\}$ for a constant θ independent of v. Defining $\hat{A}(\mathbf{U}) = \{v : T_v(\mathbf{U}) \leq \theta\}$, the Bonferroni upper-bound is

$$\mathrm{FWER} \leq \sum_{v \in A^c} \mathbb{P}(T_v(\mathbf{U}) \leq \theta) \leq |V| \max_{v \in A^c} \mathbb{P}(T_v(\mathbf{U}) \leq \theta).$$

To ensure that FWER $\leq \alpha$, $\theta = \theta_B$ is selected such that $\mathbb{P}(T_v(\mathbf{U}) \leq \theta_B) \leq \alpha/|V|$ whenever $v \in A^c$. The Bonferroni bound can only be marginally improved (see, in particular the estimator in,[4] which will be referred to as Bonferroni-Holm in the rest of the thesis) in the general case. The Bonferroni-Holm algorithm can be described as follows. Assume that under $H_0(v)$, the statistic $T_v(\mathbf{U})$ follows (dominates) a uniform distribution on $[0, 1]$. First, order the statistics $(T_v(\mathbf{U}), v \in V)$ in an ascending order. Denote the corresponding ordered statistics by $T_{(1)}(\mathbf{U}), T_{(2)}(\mathbf{U}), ...T_{(|V|)}(\mathbf{U})$ and the corresponding ranks of the T_v's by $r(v)$ for each v. Compute $k^*(\mathbf{U})$ as $\min\{k : T_{(k)}(\mathbf{U}) > \frac{\alpha}{|V|-k+1}\}$. Given $k^*(\mathbf{U})$, the detection set $\hat{A}(\mathbf{U})$ is therefore defined as $\{v \in V : r(v) < k^*(\mathbf{U})\}$. This procedure is an example of a procedure where the rejection regions depend on the random vector $\mathbf{U}$. While alternative procedures (including permutation tests) can be designed to take advantage of correlations among tests, the Bonferroni bound is sharp when $|V| \gg |A|$ and tests are independent. We will postpone a brief description of a class of resampling methods taking advantage of the correlations between tests to Section 1.6.

The second criterion is the false discovery rate. Before defining this criterion, let us mention that it is always smaller than the FWER. Therefore, any procedure controlling the FWER will control the false discovery rate.

FDR: Several procedures in the literature[7–11] also focused on controlling the false discovery rate (FDR), which is the expected ratio between the number of false alarms

$|A^c \cap \hat{A}(\mathbf{U})|$ and the number of discoveries $|\hat{A}(\mathbf{U})|$:

$$\mathbb{E}\left(\frac{|\hat{A}(\mathbf{U}) \cap A^c|}{|\hat{A}(\mathbf{U})| \vee 1}\right).$$

The most famous procedure is certainly the Benjamini-Hochberg[7] procedure. Here also, we assume that the statistics $T_v(\mathbf{U})$ are following (dominating) a uniform distribution on $[0,1]$. Moreover, we assume that $(T_v(\mathbf{U}), v \in V)$ are independent [3]. The procedure consists in computing $k^*(\mathbf{U}) = \max\{k : T_{(k)}(\mathbf{U}) \leq \frac{\alpha k}{|V|}\}$. Then define $\hat{A}(\mathbf{U})$ as $\{v \in V : r(v) \leq k^*(\mathbf{U})\}$. One can view the Benjamini-Hochberg method in the following equivalent way. If we fix a threshold t and define the detection set $\hat{A}_t(\mathbf{U})$ as all the indices having a corresponding statistic less than t, a reasonable estimator of the FDR will be $\frac{|V|t}{|\hat{A}_t(\mathbf{U})| \vee 1}$. Notice that in this expression, $|V|$ can be viewed as a conservative estimator of A^c. The Benjamini-Hochberg procedure computes $t^*(\mathbf{U})$ as $\max\{t : \frac{|V|t}{|\hat{A}_t(\mathbf{U})| \vee 1} \leq \alpha\}$. In the same spirit and under the same conditions,[11] propose a family of estimators $|\hat{A}^c(\mathbf{U}, \lambda)|$ of $|A^c|$ indexed by a positive real number λ between 0 and 1. For each λ, the estimator is defined as: .

$$|\hat{A}^c(\mathbf{U}, \lambda)| = \frac{|\{v : T_v(\mathbf{U}) \geq \lambda\}| + 1}{1 - \lambda}.$$

The idea behind this estimator relies on two observations. Under the null hypothesis, the statistic $T_v(\mathbf{U})$ will follow a uniform on $[0,1]$. If we choose λ large enough,

[3] In fact, this assumption can be relaxed to the so called positive regression dependency assumption. This assumption will be defined precisely later in this thesis

all the statistics $T_v(\mathbf{U})$ where $H_0(v)$ is false will be less than λ. With this estimator, a threshold $t^*(\mathbf{U}, \lambda)$ is defined as $\max\{t \leq \lambda : \frac{|\hat{A}^c(\mathbf{U},\lambda)|t}{|\hat{A}_t(\mathbf{U})|\vee 1} \leq \alpha\}$, and the set $\hat{A}(\mathbf{U}, \lambda)$ is defined as the set $\{v \in V : T_v(\mathbf{U}) \leq t^*(\mathbf{U}, \lambda)\}$. In reality, the described procedure controls the FDR at the desired level α for any deterministic value of λ. However, the author proposes a heuristic for choosing the value of λ based on Bootstrap methods.[12–14] It is important to note that with such a λ, there is no theoretical guarantee to control the FDR at the desired level. On another note,[10] proposes a generalization of the Benjamini-Hochberg method that allows to control the FDR for any dependency structure between the tests, at the expense of the power of detections. It replaces the definition of $k^*(\mathbf{U})$ in Benjamini-Hochberg by $k^*(\mathbf{U}) = \max\{k : T_{(k)}(\mathbf{U}) \leq \frac{\alpha\beta(k)}{|V|}\}$, where $\beta(k)$ is any function having the form $\beta(k) = \sum_{r=0}^{k} r\nu(r)$ and ν is any probability measure on the set of integers. In the rest of the thesis, we will refer to the family of these procedures as ν-procedures. To compare, with the Benjamini-Hochberg procedure, where $\beta(k)$ is replaced by k, one can immediately see that $\beta(k)$ is always less or equal than k , with equality at a certain integer k_0 only if $\nu(k)$ is a Dirac measure on k_0. Therefore, these methods will always be more conservative. This family of procedures encompasses different procedures that have been proposed prior to.[10] We refer for example to.[8,15] Recently,[16–18] proposed the so called knock-off filters method that we briefly describe here.

Assume that the random vector $\mathbf{U}$ has the form $(Y, (X_v), v \in V)$, where Y and the X_v's are random vectors. For each v, the hypothesis $H_0(v)$ is: Y and X_v are indepen-

dent given $(X_{v'}, v' \in V \setminus \{v\})$. The procedure requires the generation of "knock-off" variables $(\tilde{X}_v, v \in V)$ satisfying the following criteria:

- $\odot$ For any subset $\mathbf{S}$ of V, the joint distribution of $(X_v, v \in V, \tilde{X}_v, v \in V)$ is the same as the distribution of the vector $\text{Swap}_{\mathbf{S}}\Big((X_v, v \in V, \tilde{X}_v, v \in V)\Big)$, where $\text{Swap}_{\mathbf{S}}\Big((X_v, v \in V, \tilde{X}_v, v \in V)\Big)$ is the vector obtained from $(X_v, v \in V, \tilde{X}_v, v \in V)$ by exchanging $(X_v, v \in \mathbf{S})$ and $(\tilde{X}_v, v \in \mathbf{S})$.

- $\odot$ The vectors $(\tilde{X}_v, v \in V)$ and Y are independent given $(X_v, v \in V)$.

Once the knock-off variables are constructed, the next step is to compute statistics $W_v\Big(Y, (X_v, v \in V, \tilde{X}_v, v \in V)\Big)$ for each v, satisfying the two following properties.

- $\odot$ $W_v\Big(Y, (X_v, v \in V, \tilde{X}_v, v \in V)\Big) = W_v\Big(Y, \text{Swap}_{\mathbf{S}}\Big(X_v, v \in V, \tilde{X}_v, v \in V\Big)\Big)$ if $v \notin \mathbf{S}$.

- $\odot$ $W_v\Big(Y, (X_v, v \in V, \tilde{X}_v, v \in V)\Big) = -W_v\Big(Y, \text{Swap}_{\mathbf{S}}\Big(X_v, v \in V, \tilde{X}_v, v \in V\Big)\Big)$ if $v \in \mathbf{S}$.

Larger values of W_v's correspond to stronger evidence to reject the null $H_0(v)$. The idea behind generating such statistics is to reduce the problem to the independent tests case, starting with any dependence structure. Finally, the detection set $\hat{A}(\mathbf{U})$ is defined in this way:

- $\odot$ Compute $t^*(\mathbf{U}) = \min\{t > 0 : \frac{1+|v \in V : W_v \leq -t|}{|v \in V : W_v \geq t|}\}$.

$\odot$ $\hat{A}(\mathbf{U}) = \{v \in V : W_v \geq t^*(\mathbf{U})\}$.

In fact, the coarse-to-fine procedures that we will propose for controlling the FDR are a modified version of (generalized) Benjamini-Hochberg methods. Even if for a future work, we intend to propose a coarse-to-fine version in the same spirit of knock-off filters, this is the only place in the thesis where we mention the knock-off filters procedure.

Other criteria have been proposed in order to generalize either the FWER or the FDR. We refer for example to.[19,20] One criterion worth mentioning is:

$$\mathbf{pFDR} = \mathbb{E}\left(\frac{|\hat{A}(\mathbf{U}) \cap A^c|}{|\hat{A}(\mathbf{U})|} \,\middle|\, |\hat{A}(\mathbf{U})| > 0\right).$$

In reality, this criterion fits very nicely in the Bayesian multiple testing framework and has a very neat interpretation. Our work however does not overlap with the Bayesian vision of multiple testing problems. We refer the interested readers to.[21,22]

Finally, let us mention that in its largest part, our work will focus on procedures controlling the first criterion, namely the FWER.

1.4 Motivation of the coarse-to-fine procedure via FWER control

In many cases, including the settings in computational biology that directly motivate this work, we find $|A| \ll |V|$, $n \ll d$ as well as small "effect sizes." This is the case, for example, in genome-wide association studies (GWAS) where $\mathbf{U} = (Y, X_v, v \in V)$ and the dependence of the "phenotype" Y on the "genotype" $(X_v, v \in V)$ is often assumed to be linear; the active set A are those v with non-zero coefficients and effect size refers to the fraction of the total variance of Y explained by a particular X_v. Under these challenging circumstances, the methods used to guarantee the FWER criterion are usually very conservative and power is limited; that is, number of true positive detections is often very small (if not null) compared to $|A|$ (the "missing heritability"). This is why the less conservative FDR criterion is sometimes preferred: it allows for a higher number of true detections, but of course at the expense of false positives. However, there are situations, such as GWAS, in which this tradeoff is unacceptable; for example, collecting more data and doing follow-up experiments may be too labor intensive or expensive, and therefore having even one false discovery may be deemed undesirable.

Clearly some additional assumptions or domain-specific knowledge are necessary to ameliorate the reduction in power resulting from controlling the FWER. Motivated by applications in genomics, we suppose the set V has a natural hierarchical structure.

In principle, it should then be possible to gain power if the active hypotheses are not randomly distributed throughout V but rather have a tendency to cluster within cells of the hierarchy. In our work , we shall consider in detail the simplest example consisting of only two levels corresponding to individual hypotheses indexed by $v \in V$ and a partition of V into non-overlapping subsets ($g \subset V, g \in G$), which we call "cells."

We propose different multiple testing strategies that are coarse-to-fine with respect to this structure. The greatest part of the work focuses on strategies controlling the FWER, and whose power will exceed standard approaches for typical models and realistic parameters when a minimal degree of clustering is present. We will also present coarse-to-fine procedures controlling the FDR. It is important to note that the clustering property is not a condition for a correct control of the FWER at a given level using our coarse-to-fine procedure, but only for its increased efficiency in discovering active hypotheses.

1.5 The coarse-to-fine paradigm

We are now ready to introduce the general pattern of our coarse-to-fine procedures. Our estimate of A will be based on two families of test statistics: $\{T_v(\mathbf{U}), v \in V\}$, as above, and $\{T_g(\mathbf{U}), g \in G\}$. Without loss of generality, assume that rejection regions will correspond to small values of these statistics. The cell-level test T_g is designed to assume small values only when g is "active," meaning that $g \cap A \neq \emptyset$. Our estimator

of A is now

$$\hat{A}(\mathbf{U}) = \{v : T_g(\mathbf{U}) \leq \theta_G,\, T_v(\mathbf{U}) \leq \theta_V\},$$

where θ_V and θ_G can possibly be random thresholds depending on the random vector $\mathbf{U}$. In other words, the index needs to pass two levels in order to reject the null hypothesis $H_0(v)$. One theoretical challenge of this method is to derive a tractable method for controlling a specified error (FWER or FDR) at a given level α. Roughly speaking, the general behavior of the coarse-to-fine proposed procedures will be the following. Given a parameter J that is the number of indices belonging to active cells [4] we compute thresholds θ_V and θ_G and then define the detection set $\hat{A}(\mathbf{U}, \theta_G, \theta_V, J)$. The thresholds θ_G and θ_V will be chosen such that our error criterion associated to the detection set will be controlled at a given level. The smaller is J, the less conservative will be the thresholds θ_G and θ_V for the same fixed level. Therefore, the more clustered the active indices, the more detections we make. While J is not known in general, we will estimate an upper bound, $\hat{J}$, based on the random vector $(\mathbf{U})$. The parameter J measures the degree of clustering of active indices inside active cells.

In contrast to the usual single level based procedures, the main technical difficulty arises from the correlation between the test statistics T_g and T_v for $v \in g$. This must be taken into account since it increases the likelihood of an individual index v

[4] Not all the procedures will be explicitly based on this parameter, but the general pattern of all the procedures is similar to procedures based on J.

being falsely declared active when the cell $g(v)$ that contains it is falsely discovered (survivorship bias). More specifically, we require estimates of quadrant probabilities under the *joint distribution* of $T_{g(v)}(\mathbf{U})$ and $T_v(\mathbf{U})$ when $g(v)$, the cell containing v, is inactive. Clearly, these probabilities will depend on the behavior of the unknown probability $\mathbb{P}$ when it belongs to subsets of the form $\cap_{v\in A_0}\mathcal{P}_v$ where A_0 is some subset of A^c. In this context, we now present the framework associated to the null hypotheses treated in this work.

1.6 Invariance of distributions under group actions

Before presenting a formal definition, let us go back to our example of pairwise comparisons between N normal populations with the same variance, and assume for simplicity that all the populations have the same size denoted by n. Let us take a family of tuples $(\{i_1, j_1\}, \{i_2, j_2\}, ...\{i_K, j_K\})$ such that $\{i_k, j_k\}$ is in A^c for every $k \leq K$. One can therefore notice that for every permutation ξ in the group $\mathfrak{S}_{2n}$ of permutations over $2n$ elements and every k, the distribution of $\xi \odot (X_{i_k,1},..., X_{i_k,n}, X_{j_k,1}, ..., X_{j_k,n})$, where $\xi \odot \mathbf{Z}$ is the vector obtained from $\mathbf{Z} \in \mathbb{R}^{\mathbf{2n}}$ by permuting its $2n$ coordinates , is invariant in that we mean that it has the same distribution for every $\xi \in \mathfrak{S}_{2n}$. Now, if we denote by $\xi \odot \Big((X_{i_k,l})_{1\leq l\leq n}, (X_{j_k,l})_{1\leq l\leq n}\Big)_{1\leq k\leq K}$

the vector $\Big(\xi \odot ((X_{i_k,l})_{1\leq l\leq n}, (X_{j_k,l})_{1\leq l\leq n})\Big)_{1\leq k\leq K}$, we notice that since $\{i_k, j_k\}$ is in A^c for every $k \leq K$, the distribution of the vector $\xi \odot \Big((X_{i_k,l})_{1\leq l\leq n}, (X_{j_k,l})_{1\leq l\leq n}\Big)_{1\leq k\leq K}$ is also invariant under the action of ξ, the group acting on the dataset being $\mathfrak{S}_{2n}$ in this case. Another example is when the random vector $\mathbf{U}$ has the form of i.i.d random variables $\big((Y^k, X^k), k = 1, ..., n)\big)$ where the Y's are real valued and the variables $X^k = (X^k_v, v \in V)$ is a high-dimensional family of variables indexed by the set V. In this setting, if $H_0(v)$ corresponds to the hypothesis that $(X^k_v, k = 1, ...n)$ is independent from $\big(Y^k, , k = 1, ..., n\big)$, then for any element ξ belonging to the group of permutations over n elements, the distribution of:

$$\xi \odot \big(Y^k, X^k_v, k = 1, ..., n; v \in A_0\big) = \big(Y^{\xi_k}, X^k_v, k = 1, ..., n; v \in A_0\big)$$

is the same for every $\xi \in \mathfrak{S}_n$ whenever the set $A_0 \subset A^c$. In fact, the elements ξ are not necessarily elements of the permutation group. Imagine for example a situation where $\mathbf{U} = (X^k_v, k = 1, ..., n; v \in V)$, where for each v the vector $(X^k_v, k = 1, ...n)$ consists of i.i.d real valued random variables. Imagine moreover that the vectors $(X^k_{v_1}, k = 1, ...n)$ and $(X^k_{v_2}, k = 1, ...n)$ are independent whenever v_1 and v_2 are different. Suppose that $H_0(v)$ for each $v \in V$ is that the distribution of $(X^k_v, k = 1, ...n)$ is symmetric. Then, if $\xi = (\xi_1, ..., \xi_n)$ is any element belonging to the group $\{-1, 1\}^n$, the distribution of:

$$\xi \odot (X^k_v, k = 1, ..., n \; and \; v \in A_0) = (\xi_k X^k_v, k = 1, ..., n \; and \; v \in A_0)$$

is the same for any ξ, whenever $A_0 \subset A^c$.

More generally, in the spirit of the previous examples, we will assume in this work that the statistical model $(\mathcal{U}, \mathcal{F}, \mathbb{P}_{(\mathbb{P}\in\mathcal{P})})$ and its associated family of null hypotheses $(H_0(v), v \in V)$ satisfy the following property. There exists a group $\mathfrak{S}$ such that for every subset A_0 of A^c, there exists a random vector $\mathbf{U}_{A_0}$ that is a measurable transformation of the random vector $\mathbf{U}$, belonging to some set $\mathcal{X}$, possibly depending on A_0, and a group action from $\odot : \mathfrak{S} \times \mathcal{X} \to \mathcal{X}$ such that the distribution of $\xi \odot \mathbf{U}_{A_0}$ is the same for every $\xi \in \mathfrak{S}$. We will say that the distribution of $\mathbf{U}_{A_0}$ is invariant under the action of $\xi \in \mathfrak{S}$.

As mentioned in Section 1.3, we are now ready to describe a procedure controlling the FWER that accounts for the dependency structure of the tests and can improve upon the Bonferroni-Holm method when the test statistics are highly correlated. This procedure plays an important role in our work in the sense that we will derive a coarse-to-fine method in the same spirit. Namely, we would like to estimate the distribution of $\min\{T_v(\mathbf{U}), v \in A^c\}$. If we knew $t^* = \max\{t : \mathbb{P}\left(\min\{T_v(\mathbf{U}), v \in A^c\} < t\right) \leq \alpha\}$, then immediately, by defining:

$$\hat{A}(\mathbf{U}) = \{v \in V : T_v(\mathbf{U}) < t^*\},$$

we would have a procedure controlling the FWER at a desired level α. Obviously, the distribution of $\min\{T_v(\mathbf{U}), v \in A^c\}$ is unknown and it is not possible to compute

t^*. To address this issue, the group action framework allows to have a conservative estimator of t^* that can be defined as following. Imagine for simplicity that we are able to generate in a short time all the elements $\xi \in \mathfrak{S}$ (we will later refer to this assumption as asymptotic resampling) and therefore have access to the distribution of $\min\{T_v(\xi \odot \mathbf{U}), v \in V\}$ given $\mathbf{U}$. Here we view the elements ξ as random elements generated uniformly on $\mathfrak{S}$. This point of view will be formalized and described precisely later in the thesis. Denote by μ the distribution associated to the ξ's. An estimator of t^* will therefore be:

$$\hat{t}(\mathbf{U}) = \max\{t : \mu\left(\xi : \min\{T_v(\xi \odot \mathbf{U}), v \in V\} < t\right) \leq \alpha\}.$$

One can simply imagine that the elements ξ as permutations. $\mu\left(\xi : \min\{T_v(\xi \odot \mathbf{U}), v \in V\} < t\right)$ is therefore the number of permutations among the total number of permutations where we observed a statistic (among all statistics) less than t. If we define $\hat{A}(\mathbf{U})$ as :

$$\hat{A}(\mathbf{U}) = \{v \in V : T_v(U) \leq \hat{t}(\mathbf{U})\},$$

then the FWER is controlled at a desired level α. Here, the rejection set depends on the random vector $\mathbf{U}$.

1.7 Applications and related work

As indicated above, our work (and some of our notation) is inspired by statistical issues arising in GWAS[23–25] and related areas in computational genomics. In the most common version of GWAS, the "genotype" of an individual is represented by the genetic states X_v at a very large family of genomic locations $v \in V$; these variations are called single nucleotide polymorphisms or SNPs. These SNPs are used as markers of genomic regions. The majority of these SNPs don't have any biological impact. The final goal of GWAS is , for a particular study, to find the rare SNPs that are associated to a certain trait or phenotype. At this stage, one needs to carefully explain what is meant by associated. In reality, there are two types of association: the first type is called direct association as opposed to the second type: indirect association.[26] If the SNP is directly associated to the trait, the SNP is the functional SNP, in the sense that it causes a change in gene expression having an effect on the particular phenotype. This can happen for example via modifications in amino acids, mRNA transcript stability or changes to transcription factor binding affinity.[27] Indirect association occurs in the situation where there is a correlation between the phenotype and a particular SNP, but the SNP is not causal. It happens when the SNP in question is correlated with a second SNP that is in direct association with the trait, but is not causing any functional change having an effect on the phenotype. When two SNPs are in a such situation, we say that they are in linkage disequilibrium (LD). When a set of SNPs are in mutual linkage disequilibrium, we say that they form an LD block. To

avoid collecting redundant information, GWAS studies try to select a representative SNP for each LD block. Such a SNP is called a tag SNP, since it is tagging the whole LD to which it belongs. As a consequence, a causal SNP can be absent from a study, and looking just for a causality relationship between SNPs and the trait can lead to missing a tag SNP that is associated to an important functional SNP. Therefore, our objective will be to find those SNPs $A \subset V$ "in (both types) association" with a given phenotype, for example a measurable trait Y such as height or blood pressure. The null hypothesis for a SNP v is that Y and X_v are independent random variables, and whereas $|V|$ may run into the millions, the set A of active variants is expected to be fewer than one hundred. Obviously, the set $\hat{A}$ will not represent a set of causal SNPs and further studies will be required, typically investigating the LD blocks associated to the set $\hat{A}$. The number of hypotheses tested $|V|$ corresponds to the number of tag SNP's, which is roughly the number of LD blocks. The number of LD blocks depends on the number of recombination events that occurred within a studied population. For example, the number of LD blocks in African populations is higher than the number of blocks in European population due to a larger number of recombination events. The number $|V|$ of blocks in a study involving European populations has a magnitude of 10^6.[28] For a given population studied, the power of detection of a given SNP v is driven by two factors: The minor allele frequency and the effect size. The minor allele frequency is simply the frequency of the allele that is less frequent in a population for the SNP v. As for the effect size of v, it is typically a statistic reflecting

the degree of dependency between Y and X_v. Different definitions of the effect size have been given in different contexts,[29] depending on whether Y is a quantitative or categorical trait in our context. In our case, we are interested in the case where Y can take binary values 0 or 1. This setting corresponds to what is defined as case control studies. X_v can take 3 possible values 0, 1 or 2. These correspond to counts of the number of minor alleles at the particular locus. In this particular setting, the effect size associated to SNP is defined as the couple of odds ratios:

$$\text{odds ratio} = \frac{P(Y=1|X_v=1,X_{v'})/P(Y=0|X_v=1,X_{v'})}{P(Y=1|X_v=0,X_{v'})/P(Y=0|X_v=0,X_{v'})}$$

and

$$\text{odds ratio} = \frac{P(Y=1|X_v=2,X_{v'})/P(Y=0|X_v=2,X_{v'})}{P(Y=1|X_v=0,X_{v'})/P(Y=0|X_v=0,X_{v'})}$$

If moreover, we assume that the second odds ratio is the square of the first ratio, then the effect size is simply defined as the first odds ratio. Figure 1.1 gives an idea about effect sizes and minor allele frequencies of certain traits.

Finally, the situation in GWAS can be well adapted to coarse-to-fine strategies because of the following reason. If the considered variants are confined to coding regions, then the set of genes provides a natural partition of V. In reality this is not a necessary condition. Instead, what is needed is any mapping/function that associates SNPs to sets of SNPs based on some biological knowledge. Also, the fact that genes are organized into pathways provides a natural three-level hierarchy.[30]

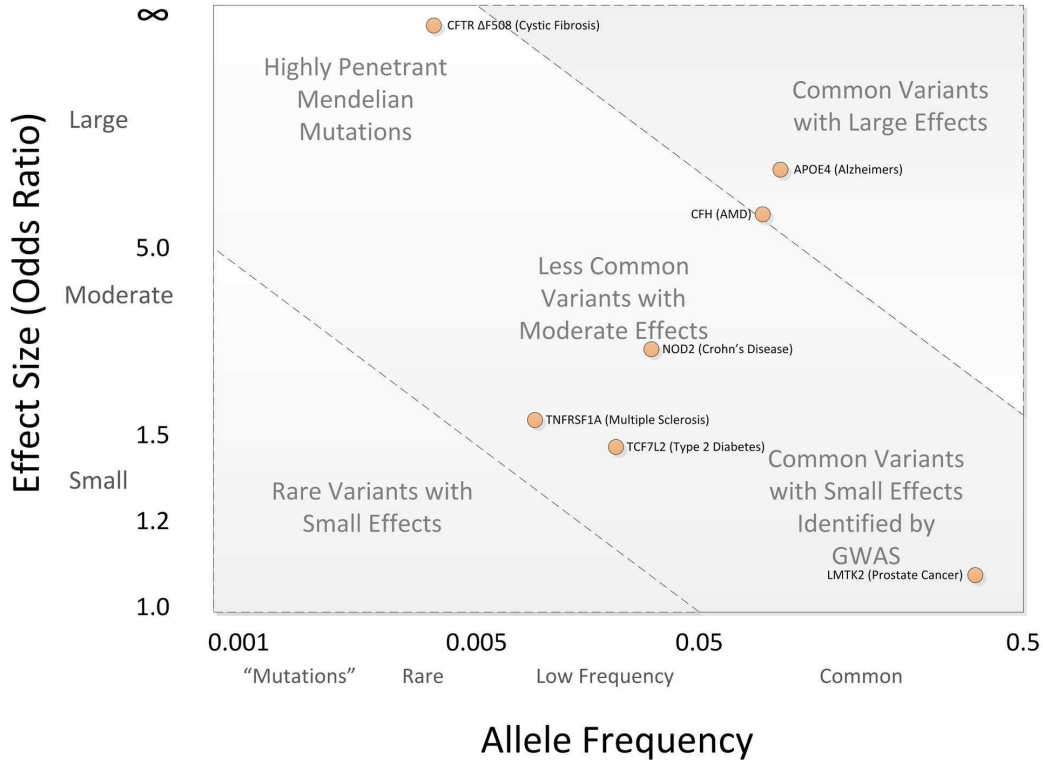

Figure 1.1: Representation of minor allele frequencies and effect sizes for some diseases. The majority of diseases are caused by common SNPs with small effect sizes. Figure taken from[1]

Another application of large-scale multiple testing is variable filtering in high-dimensional prediction: the objective is to predict a categorical or continuous variable Y based on a family of potentially discriminating features $X_v, v \in V$. Learning a predictor $\hat{Y}$ from i.i.d. samples of $U = (Y, X_v, v \in V)$ is often facilitated by limiting *a priori* the set of features utilized in training $\hat{Y}$ to a subset $A \subset V$ determined by testing the features one-by-one for dependence on Y and setting a significance threshold. In most applications of machine learning to artificial perception, no premium is placed on pruning A to a highly distinguished subset; indeed, the particular set of selected features is rarely examined or considered of significance. In contrast, the identities of the particular features selected and appearing in decision rules are often of keen interest in computational genomics, e.g., discovering cancer biomarkers, where the variables X_v represent "omics" data (e.g., gene expression), and Y codes for two possible cellular or disease phenotypes. Obtaining a "signature" $\hat{A}$ devoid of false positives can be beneficial in understanding the underlying biology and interpreting the decision rules. In this case the Gene Ontology (GO)[31] provides a very rich hierarchical structure, one example being the organization of genes in pathways. Indeed, building predictors to separate "driver mutations" from "passenger mutations" in cancer would appear to be a promising candidate for coarse-to-fine testing due to the fact that drivers are known to cluster in pathways.

There is a literature on coarse-to-fine pattern recognition (see, e.g.,[32] and the references therein), but the emphasis has traditionally been on computational efficiency

rather than error control. Computation is not considered here. Moreover, in most of this work, especially applications to vision and speech, the emphasis is on detecting true positives (e.g., patterns of interest such as faces) at the expense of false positives. Simply "reversing" the role of true positives and negatives is not feasible due to the loss of reasonable invariance assumptions; in effect, every pattern of interest is unique.

In,[33] a hierarchical testing approach is used in the context of the FWER. However, the intention in this paper is to improve the power of detection relative to the Bonferroni-Holm methods only at level of clusters of hypotheses; in contrast to our method, the two approaches have comparable power at the level of individual hypotheses. To briefly describe the method proposed in the context of two levels, the procedure consists in computing three types of statistics. First, compute a global statistic $T_V(\mathbf{U})$ in order to reject or not the global null hypothesis: For every $v \in V$ $H_0(v)$ is true. Under the global null the statistic is assumed to follow (dominates) a uniform distribution on $[0, 1]$. At a second stage, compute the statistics $T_g(\mathbf{U})$ for every cell g. For each g, if $H_0(v)$ is true for every $v \in g$, then $T_g(\mathbf{U})$ would follow (dominate) a uniform distribution on $[0, 1]$. Finally, compute the statistics $T_v(\mathbf{U})$ at the indices level v, assuming also that these statistics are following (dominating) a uniform distribution on $[0, 1]$. Once the statistics are computed, the method first decides to reject the global null if and only if $T_V(\mathbf{U}) \leq \alpha$. If the global null is rejected, reject the null for each cell g if and only if $T_g(\mathbf{U}) \leq \frac{|g|\alpha}{|V|}$. Finally, each null $H_0(v)$ at the indices level is rejected if and only if $T_{g(v)}(\mathbf{U}) \leq \frac{|g|\alpha}{|V|}$ and $T_v(\mathbf{U}) \leq \frac{\alpha}{|V|}$, where

$g(v)$ is the cell containing v for each $v \in V$. One immediately notices that there is no improvement over the Bonferroni bound at the indices level. In reality, the interest of the mentioned work consists in obtaining a Bonferroni bound at the cell level, but controlling an error that is more conservative than the FWER. The point of view that is taken in[33] is the following. We are not just interested in a detection set $\hat{A}(\mathbf{U})$ at the indices level. The final detection set that is considered comes in three stages: $\hat{A}_{global}(\mathbf{U})$, $\hat{A}_G(\mathbf{U})$ and $\hat{A}_V(\mathbf{U})$. The first detection set is either the empty set or a singleton and corresponds to rejecting or not the global null. The second detection consists in all the cells where the null has been rejected. Finally, the third detection corresponds to the same detection set that we are considering in our work and consists in all the indices where the null has been rejected. In fact, the procedure that is described here controls an error that is more conservative than the FWER that is usually considered: $\mathbb{P}\left(\hat{A}_V(\mathbf{U}) \cap A_V^c \neq \emptyset\right)$, where A_V^c is the true set of non active indices. Let us denote by A_{global}^c the set that is either the empty set if the global null is false or the singleton V otherwise. Also, denote by A_G^c the true set of non-active cells. The FWER that is controlled here is then the following:

$$\mathbb{P}\left(\{\hat{A}_{global}(\mathbf{U}) \cap A_{global}^c \neq \emptyset\} \bigcup \{\hat{A}_G(\mathbf{U}) \cap A_G^c \neq \emptyset\} \bigcup \{\hat{A}_V(\mathbf{U}) \cap A_V^c \neq \emptyset\}\right).$$

It is indeed guaranteed to be less than α in the described method. To compare with our procedures, we are not interested in controlled errors at the coarse level and our

goal is to control only the error at the indices level, and at the same time improve the Bonferroni bounds. One needs to mention though that a slight improvement over the Bonferroni bound is derived in the mentioned paper, in the particular case where the cells g form a partition of V of size 2. In this case, the detected indices are those such that $T_{g(v)}(\mathbf{U}) \leq \frac{|g|\alpha}{|V|}$ and $T_v(\mathbf{U}) \leq \frac{2\alpha}{|V|}$.

Finally, two-stage procedures have been proposed (see for example[15,34]) to control the FDR. However, these procedures are coarse-to-fine in the sense that the two stages for each v are based on the statistics $T_v(\mathbf{U})$. The general pattern of these strategies consists in computing a first stage threshold on the statistics $(T_v(\mathbf{U}), v \in V)$ that will output a first detection set. At a second stage a second threshold is computed depending on the first detection set and threshold.

1.8 Organization of the thesis

In the next chapter, we will present a procedure considered as a Bonferroni coarse-to-fine version, in the sense that it is derived using Bonferroni bounds. This procedure will control the FWER. In Chapter 4, we will take into account the setting where the statistics associated to the multiple hypotheses are correlated and derive a coarse-to-fine procedure adapted for high correlations within the statistics associated to the hypotheses. This procedure will also control the FWER. Finally, in Chapter 5 we will consider a procedure controlling the FDR. We will first consider coarse-to-

fine procedures controlling the FDR under dependency structure assumptions within the statistics associates to the hypotheses, and then present procedures controlling the FDR under any dependency structure at the expense of the power of detection compared to the first family of procedures.

Chapter 2

A Bonferroni coarse-to-fine procedure

This chapter is organized in the following way. We will start by a formal introduction of the coarse-to-fine framework. This framework will be used for the remaining part of the manuscript. We then give a Bonferroni coarse-to-fine method in the context described in section 1.6 , assuming asymptotic resampling. The next section provides a modification of the former procedure that takes into account finite resampling. Finally, in order to illustrate in a simpler way the ideas behind the Bonferroni coarse-to-fine procedure, we propose a model based version of the method, where we assume a gaussian linear model. The performance of the methods that are derived in this chapter are compared at the end of the chapter via simulations.

2.1 Coarse-to-fine framework

The finite family of null hypotheses will be denoted by $(H_0(v), v \in V)$, where H_0 is either true or false. We are interested in the active set of indices, $A = \{v \in V : H_0(v) = \text{false}\}$ and will write $V_0 = A^c$ for the set of inactive indices. Suppose our data $\mathbf{U}$ takes values in $\mathcal{U}$. The set $\hat{A}(\mathbf{U})$ is commonly designed based on individual rejection regions $\Gamma_v \subset \mathcal{U}$, with $\hat{A}(\mathbf{U}) = \{v : \mathbf{U} \in \Gamma_v\}$. As indicated in the previous section, in the conservative Bonferroni approach, the FWER is controlled at level α by assuming $|V| \max_{v \in V_0} \mathbb{P}(\mathbf{U} \in \Gamma_v) \leq \alpha$. If the rejection regions are designed so that this probability is independent of v whenever $H_0(v)$ is true, then the condition boils down to $\mathbb{P}(\mathbf{U} \in \Gamma_v) \leq \alpha/|V|$ for $v \in V_0$. Generally, $\Gamma_v = \{u \in \mathcal{U} : T_v(u) \leq t\}$ for a constant t (or t_v) for some family of test statistics $(T_v, v \in V)$.

While there is not much to do in the general case to improve on the Bonferroni method, it is possible to improve power if V is structured and one has prior knowledge about the way the active hypotheses are organized relative to this structure. In this paper, we consider a coarse-to-fine framework in which V is provided with a partition G, so that $V = \bigcup_{g \in G} g$, where the subsets $g \subset V$ (which we will call cells) are non-overlapping. For $v \in V$, we let $g(v)$ denote the unique cell g that contains it. The "coarse" step selects cells likely to contain active indices, followed by a "fine" step in which a Bonferroni or equivalent procedure is applied only to hypotheses included in the selected cells. More explicitly, we will associate a rejection region Γ_g to each

$g \in G$ and consider the discovery set

$$\hat{A}(\mathbf{U}) = \{v \in V : \mathbf{U} \in \Gamma_{g(v)} \cap \Gamma_v\}. \tag{2.1}$$

We will say that a cell g is active if and only if $g \cap A \neq \emptyset$, which we shall also express as $H_0(g) =$ false, implicitly defining $H_0(g)$ as the logical "and" of all $H_0(v), v \in g$. We will also consider the double null hypothesis $H_{00}(v) = H_0(g(v))$ of v belonging in an inactive cell (which obviously implies that v is inactive too), and we will let $V_{00} \subset V_0$ be the set of such v's.

Let $|g|$ denote the size of each cell g in G, G_0 and G_0^c respectively the set of non-active cells and active cells. Then, define $J = \sum_{g \in G_0^c} |g|$, the number of active indices contained in active cells. We will develop our procedure under the assumption that J is known, or, at least bounded from above. While this can actually be a plausible assumption in practice, we will relax it in section 2.2 in which we will design a procedure to estimate a bound on J.

With this notation, we have the following result:

Proposition 2.1.1. *With $\hat{A}$ defined by* (2.1)*:*

$$\mathrm{FWER}(\hat{A}) \leq |V| \max_{v \in V_{00}} \mathbb{P}\left(\mathbf{U} \in \Gamma_{g(v)} \cap \Gamma_v\right) + J \max_{v \in V_0} \mathbb{P}\left(\mathbf{U} \in \Gamma_v\right).$$

Notice that the result will obviously still be valid if we replace J by an upper bound.

Proof. This is just the Bonferroni bound applied to the decomposition

$$\begin{aligned}(\hat{A}(\mathbf{U}) \cap V_0 \neq \emptyset) &= \bigcup_{v \in V_{00}} (\mathbf{U} \in \Gamma_{g(v)} \cap \Gamma_v) \cup \bigcup_{v \in V_0 \setminus V_{00}} (\mathbf{U} \in \Gamma_{g(v)} \cap \Gamma_v) \\ &\subset \bigcup_{v \in V_{00}} (\mathbf{U} \in \Gamma_{g(v)} \cap \Gamma_v) \cup \bigcup_{v \in V_0 \setminus V_{00}} (\mathbf{U} \in \Gamma_v)\end{aligned}$$

so that

$$P(\hat{A}(\mathbf{U}) \cap V_0 \neq \emptyset) \leq |V_{00}| \max_{v \in V_{00}} \mathbb{P}\left(\mathbf{U} \in \Gamma_{g(v)} \cap \Gamma_v\right) + |V_0 \setminus V_{00}| \max_{v \in V_0} \mathbb{P}(\mathbf{U} \in \Gamma_v)$$

and the proposition results from $|V_{00}| \leq |V|$ and $|V_0 \setminus V_{00}| \leq J$. □

The sets Γ_g and Γ_v will be designed using statistics $T_g(\mathbf{U})$ and $T_v(\mathbf{U})$ setting $\Gamma_g = (T_g(\mathbf{U}) \leq \theta_G)$ and $\Gamma_v = (T_v(\mathbf{U}) \leq \theta_V)$ for some constants θ_G and θ_V. Letting $p_{00}(\theta_G, \theta_V)$ be an upper-bound of $\mathbb{P}\left((T_{g(v)}(\mathbf{U}) \leq \theta_G) \cap (T_v(\mathbf{U}) \leq \theta_V)\right)$ for $v \in V_{00}$ and $p_0(\theta_V)$ of $\mathbb{P}\left(T_v(\mathbf{U}) \leq \theta_V\right)$ for $v \in V_0$, the previous upper bound becomes

$$\mathrm{FWER}(\hat{A}) \leq |V|\, p_{00}(\theta_G, \theta_V) + J\, p_0(\theta_V). \tag{2.2}$$

In the following sections our goal will be to design θ_G and θ_V such that this upper bound is smaller than a predetermined level α. Controlling the second term will lead to less conservative choices of the constant θ_V (compared to the Bonferroni estimate), as soon as $J \ll |V|$, depending on the degree of clustering, the probability p_{00} of false detection in the two-step procedure can be made much smaller than p_0 without

harming the true detection rate and the coarse-to-fine procedure will yield an increase in power for a given FWER. We require tight estimates of p_{00} and taking into account the correlation between $T_{g(v)}(\mathbf{U})$ and $T_v(\mathbf{U})$ is necessary to deal with "survivorship bias."

2.2 Non-parametric coarse-to-fine testing

2.2.1 Notation

Recall that $\mathbf{U}$ denotes the random variable representing all the data, taking values in $\mathcal{U}$. We will build our procedure from user-defined *scores*, denoted ρ_v (at the locus level) and ρ_g (at the cell level), both defined on $\mathcal{U}$, i.e., functions of the observed data.

Moreover, we assume that there exists a group action of some group $\mathfrak{S}$ on $\mathcal{U}$, which will be denoted

$$(\xi, \mathbf{u}) \mapsto \xi \odot \mathbf{u}.$$

The product in $\mathfrak{S}$ will be denoted $(\xi, \xi') \mapsto \xi\xi'$. For example, if the observation is a realization of an i.i.d. family of random variables $U = ((Y^k, X^k), k = 1, \dots n)$ where the Y's are real-valued and the variables $X^k = (X^k_v, v \in V)$ is a high-dimensional family of variables indexed by the set V, one will take $\mathbf{U} = \big((Y^k, X^k), k = 1, \dots, n\big)$.

$\mathfrak{S}$ will be the permutation group of $\{1, \ldots, n\}$ with

$$\xi \odot \mathbf{U} = \left((Y^{\xi_k}, X^k), k = 1, \ldots, n\right).$$

To simplify the discussion, we will assume that $\mathfrak{S}$ is finite and denote by μ the uniform probability measure on $\mathfrak{S}$, so that

$$\int_{\mathfrak{S}} f(\xi) d\mu(\xi) = \frac{1}{|\mathfrak{S}|} \sum_{\xi \in \mathfrak{S}} f(\xi).$$

Our running assumption will be that,

1. For any $v \in V_{00}$, the joint distribution of $(\rho_{g(v)}((\xi'\xi) \odot \mathbf{U}), \rho_v((\xi'\xi) \odot \mathbf{U}))_{\xi' \in \mathfrak{S}}$ is independent of $\xi \in \mathfrak{S}$.

2. For any $v \in V_0$, the joint distribution of $(\rho_v((\xi'\xi) \odot \mathbf{U}))_{\xi' \in \mathfrak{S}}$ is independent of $\xi \in \mathfrak{S}$.

We will also use the following well-known result.

Lemma 2.2.1. *Let X be a random variable and let $F_X(x) = P(X \leq x)$ denote its cumulative distribution function, with left limit $F_X^-(x) = P(X < x)$. Define, for $z \in [0, 1]$*

$$\bar{F}_X(x, z) := (1 - z)\left(1 - F_X^-(x)\right) + z\,(1 - F_X(x)) = \mathbb{P}(X > x) + (1 - z)\mathbb{P}(X = x).$$

Then, if $Z : \Omega \to [0, 1]$ *is uniformly distributed and independent from* X*, one has, for* $t \in [0, 1]$,

$$\mathbb{P}\left(\bar{F}_X(X, Z) \leq t\right) = t.$$

The reader can refer, for example, to[35] for a proof of this lemma in the case of discrete variable X (which will suffice for our purposes).

2.2.2 Asymptotic resampling scores

Let $Z : \Omega \to [0, 1]$ be uniformly distributed and independent of $\mathbf{U}$. We define the asymptotic scores at the cell and variable level by

$$T_g(\mathbf{U}, Z) = \mu\left(\xi : \rho_g(\mathbf{U}) < \rho_g(\xi \odot \mathbf{U})\right) + Z\ \mu\left(\xi : \rho_g(\mathbf{U}) = \rho_g(\xi \odot \mathbf{U})\right) \tag{2.3}$$

and

$$T_v(\mathbf{U}, Z) = \mu\left(\xi : \rho_v(\mathbf{U}) < \rho_v(\xi \odot \mathbf{U})\right) + Z\ \mu\left(\xi : \rho_v(\mathbf{U}) = \rho_v(\xi \odot \mathbf{U})\right) \tag{2.4}$$

$T_g(\mathbf{U}, Z)$ and $T_v(\mathbf{U}, Z)$ are the typical statistics used in permutation tests, estimating the proportion of scores that are higher than the observed one after randomizing the sample using the group action, while counting ties with a uniformly distributed weight.

For the coarse-to-fine procedure, we will need one more "conditional" statistic.

For a given constant θ_G and a uniform random variable $\tilde{Z}$ independent of $\mathbf{U}$ and Z, we define

$$N_g^{\theta_G}(\mathbf{U}, Z) = \mu\left(\xi : T_g(\xi \odot \mathbf{U}, Z) \leq \theta_G\right). \tag{2.5}$$

We then let

$$T_v^{\theta_G}(\mathbf{U}, Z, \tilde{Z}) = \frac{1}{N_{g(v)}^{\theta_G}(\mathbf{U}, Z)} \mu\left(\xi : \rho_v(\mathbf{U}) < \rho_v(\xi \odot \mathbf{U}); T_{g(v)}(\xi \odot \mathbf{U}, Z) \leq \theta_G\right) + \frac{\tilde{Z}}{N_{g(v)}^{\theta_G}(\mathbf{U}, Z)} \times \mu\left(\xi : \rho_v(\mathbf{U}) = \rho_v(\xi \odot \mathbf{U}); T_{g(v)}(\xi \odot \mathbf{U}, Z) \leq \theta_G\right). \tag{2.6}$$

We call our scores asymptotic in this section because exact expectations over μ cannot be computed in general, and can only be obtained as limits of Monte-Carlo samples. The practical finite-sample case will be handled in the next section.

With this notation, we let

$$\hat{A} = \{v : T_{g(v)}(\mathbf{U}, Z) \leq \theta_G \text{ and } T_v^{\theta_G}(\mathbf{U}, Z, \tilde{Z}) \leq \theta_V \text{ and } T_v(\mathbf{U}, Z) \leq \theta'_V\}$$

which depends on the choice of three constants, θ_V, θ_G and θ'_V. We then have:

Theorem 2.2.1. *For all $v \in V_0$:*

$$\mathbb{P}\left(v \in \hat{A}\right) \leq \theta'_V \tag{2.7}$$

and for all $v \in V_{00}$,

$$\mathbb{P}\left(v \in \hat{A}\right) \leq \theta_G \theta_V \tag{2.8}$$

This result tells us how to control the FWER for a two-level permutation test based on any scores in the (generally intractable) case in which we can exactly compute the test statistics, when we declare an index v active if and only if $T_g(\mathbf{U}, Z) \leq \theta_G$ and $T_v^{\theta_G}(\mathbf{U}, Z, \tilde{Z}) \leq \theta_V$ and $T_v(\mathbf{U}, Z) \leq \theta'_V$ (or $\max\left(T_v^{\theta_G}(\mathbf{U}, Z, \tilde{Z}), \frac{\theta_V}{\theta'_V} T_v(\mathbf{U}, Z)\right) \leq \theta_V$ if one wants to use a single v-indexed statistic as considered in section 2.1).

Proof. For (2.7), we use a standard argument justifying randomization tests, that we provide here for completeness. If $v \in V_0$, we have

$$\begin{aligned}\mathbb{P}\left(v \in \hat{A}\right) &= \mathbb{P}\left(T_g(\mathbf{U}, Z) \leq \theta_G; T_v^{\theta_G}(\mathbf{U}, Z, \tilde{Z}) \leq \theta_V; T_v(\mathbf{U}, Z) \leq \theta'_V\right) \\ &\leq \mathbb{P}\left(T_v(\mathbf{U}, Z) \leq \theta'_V\right).\end{aligned}$$

From the invariance assumption, we have

$$\begin{aligned}\mathbb{P}\left(T_v(\mathbf{U}, Z) \leq \theta'_V\right) &= \mathbb{P}\left(T_v(\xi \odot \mathbf{U}, Z) \leq \theta'_V\right) \text{ for all } \xi \in \mathfrak{S} \\ &= \int_{\mathfrak{S}} \mathbb{P}\left(T_v(\xi \odot \mathbf{U}, Z) \leq \theta'_V\right) d\mu(\xi) \\ &= \mathbb{E}\left(\mathbb{E}\left(\mu\left(\xi : T_v(\xi \odot \mathbf{U}, Z) \leq \theta'_V\right) | \mathbf{U}\right)\right)\end{aligned}$$

It now remains to remark that, for fixed $\mathbf{U}$, $T_v(\xi \odot \mathbf{U}, Z) = \bar{F}_{\zeta_{\mathbf{U}}}(\zeta_{\mathbf{U}}(\xi), Z)$ with

$\zeta_{\mathbf{U}}(\xi') = \rho_v(\xi' \odot \mathbf{U})$ for $x' \in \mathfrak{S}$. Therefore, by Lemma 2.2.1,

$$\mathbb{E}\left(\mu\left(\xi : T_v(\xi \odot \mathbf{U}, Z) \leq \theta'_V\right)|\mathbf{U}\right) = \mathbb{E}\left(\mu\left(\xi : \bar{F}_{\zeta_{\mathbf{U}}}(\zeta_{\mathbf{U}}(\xi), Z) \leq \theta'_V\right)|\mathbf{U}\right) = \theta'_V, \quad (2.9)$$

which proves (2.7). Similarly, one has

$$\mathbb{E}(N_g^{\theta_G}(\mathbf{U}, Z)|\mathbf{U}) = \mathbb{P}\left(T_g(\mathbf{U}, Z) \leq \theta_G|\mathbf{U}\right) = \theta_G. \quad (2.10)$$

Let us now prove (2.8), assuming $v \in V_{00}$ and letting $g = g(v)$. We write

$$\mathbb{P}\left(v \in \hat{A}\right) \leq \mathbb{P}\left(T_{g(v)}(\mathbf{U}, Z) \leq \theta_G; T_v^{\theta_G}(\mathbf{U}, Z, \tilde{Z}) \leq \theta_V\right).$$

and find an upper bound for the right-hand side of the inequality. Using the invariance assumption, we have

$$\begin{aligned}
&\mathbb{P}\left(T_g(\mathbf{U}, Z) \leq \theta_G; T_v^{\theta_G}(\mathbf{U}, Z) \leq \theta_V\right) \\
&= \int_{\mathfrak{S}} \mathbb{P}\left(T_g(\xi' \odot \mathbf{U}, Z) \leq \theta_G; T_v^{\theta_G}(\xi' \odot \mathbf{U}, Z, \tilde{Z}) \leq \theta_V\right) d\mu(\xi') \\
&= \mathbb{E}\left(\mu\left(\xi' : T_g(\xi' \odot \mathbf{U}, Z) \leq \theta_G; T_v^{\theta_G}(\xi' \odot \mathbf{U}, Z, \tilde{Z}) \leq \theta_V\right)\right) \\
&= \mathbb{E}\left(\mathbb{E}\left(\mu\left(\xi' : T_g(\xi' \odot \mathbf{U}, Z) \leq \theta_G; T_v^{\theta_G}(\xi' \odot \mathbf{U}, Z, \tilde{Z}) \leq \theta_V\right)|\mathbf{U}, Z\right)\right)
\end{aligned}$$

Notice that, since μ is right-invariant, we have $N_g^{\theta_G}(\xi' \odot \mathbf{U}, Z) = N_g^{\theta_G}(\mathbf{U}, Z)$ for all ξ' and

$$
\begin{aligned}
&T_v^{\theta_G}(\xi^{'} \odot \mathbf{U}, Z, \tilde{Z}) \\
&= \frac{1}{N_g^{\theta_G}(\xi' \odot \mathbf{U}, Z)} \mu\left(\xi : \rho_v(\xi^{'} \odot \mathbf{U}) < \rho_v((\xi \circ \xi^{'}) \odot \mathbf{U}); T_g((\xi \circ \xi') \odot \mathbf{U}, Z) \leq \theta_G\right) \\
&\quad + \frac{\tilde{Z}}{N_g^{\theta_G}(\xi' \odot \mathbf{U}, Z)} \mu\left(\xi : \rho_v(\xi^{'} \odot \mathbf{U}) = \rho_v((\xi \circ \xi^{'}) \odot \mathbf{U}); T_g((\xi \circ \xi') \odot \mathbf{U}, Z) \leq \theta_G\right) \\
&= \frac{1}{N_g^{\theta_G}(\mathbf{U}, Z)} \mu\left(\xi : \rho_v(\xi^{'} \odot \mathbf{U}) < \rho_v(\xi \odot \mathbf{U}); T_g(\xi \odot \mathbf{U}, Z) \leq \theta_G\right) \\
&\quad + \frac{\tilde{Z}}{N_g^{\theta_G}(\mathbf{U}, Z)} \mu\left(\xi : \rho_v(\xi^{'} \odot \mathbf{U}) = \rho_v(\xi \odot \mathbf{U}); T_g(\xi \odot \mathbf{U}, Z) \leq \theta_G\right)
\end{aligned}
$$

Let $\tilde{\mu}$ denote the probability μ conditional to the event $(T_g(\xi \odot \mathbf{U}, Z) \leq \theta_G)$ ($\mathbf{U}$, Z and $\tilde{Z}$ being fixed). Then

$$
\frac{1}{N_g^{\theta_G}(\mathbf{U}, Z)} \mu\left(T_g(\xi' \odot \mathbf{U}, Z) \leq \theta_G; T_v^{\theta_G}(\xi' \odot \mathbf{U}, Z, \tilde{Z}) \leq \theta_V\right) = \tilde{\mu}\left(\xi' : p(\xi', \tilde{Z}) \leq \theta_V\right),
$$

where

$$
p(\xi', \tilde{Z}) = \tilde{\mu}\left(\xi : \rho_v(\xi \odot \mathbf{U}) > \rho_v(\xi' \odot \mathbf{U})\right) + \tilde{Z}\tilde{\mu}\left(\xi : \rho_v(\xi \odot \mathbf{U}) = \rho_v(\xi' \odot \mathbf{U})\right)
$$

Hence, Lemma 2.2.1 implies that:

$$
\mathbb{E}\left(\frac{1}{N_g^{\theta_G}(\mathbf{U}, Z)} \mu\left(\xi' : T_g(\xi' \odot \mathbf{U}, Z) \leq \theta_G; T_v^{\theta_G}(\xi^{'} \odot \mathbf{U}, Z, \tilde{Z}) \leq \theta_V\right) \middle| \mathbf{U}, Z\right) = \theta_V.
$$

Hence,

$$\mathbb{P}\left(T_g(\mathbf{U}, Z) \leq \theta_G; T_v^{\theta_G}(\mathbf{U}, Z, \tilde{Z}) \leq \theta_V\right)$$
$$= \mathbb{E}\left(N_g^{\theta_G}(\mathbf{U}, Z)\theta_V\right) = \theta_V \mathbb{E}\left(N_g^{\theta_G}(\mathbf{U}, Z)\right) = \theta_V \theta_G.$$

□

Note that Theorem 2.2.1 is still true if one takes $Z = \tilde{Z} = 1$ in the definition of the test statistics, because the obtained detection set would then be a subset of $\hat{A}$. This would have resulted in a simpler expression in which ties are fully counted, with very little practical loss because the probability of ties in over such permutations is typically minuscule. However, equality in equations such as (2.9) will be needed in the proof of Theorem 2.2.2.

As an immediate corollary, we have:

Corollary 2.2.1.

$$\text{FWER}(\hat{A}) \leq |V|\theta_G\theta_V + J\theta'_V.$$

As mentioned above, this result does not have practical interest because it requires applying all possible permutations to the data. In practice, a random subset of permutations is picked instead, and we develop the related theory in the next section.

2.2.3 Finite resampling scores

We now replace T_g, $T_v^{\theta_G}$ and T_v with Monte-Carlo estimates and describe how the upper bounds in Theorem 2.2.1 need to be modified. We therefore introduce an i.i.d. random sample $\boldsymbol{\xi} = (\xi_1, ..., \xi_K) : \Omega \to \mathfrak{S}^K$, where $\xi_k \sim \mu$ and K is a positive integer. We also introduce the empirical measure:

$$\hat{\mu}_{\boldsymbol{\xi}} = \frac{1}{K} \sum_{k=1}^{K} \delta_{\xi_k}.$$

With this notation, we let:

$$\hat{T}_g(\mathbf{U}, \boldsymbol{\xi}) = \hat{\mu}_{\boldsymbol{\xi}} \left(\xi' : \rho_g(\mathbf{U}) \leq \rho_g(\xi' \odot \mathbf{U}) \right),$$

$$\check{T}_g^-(\mathbf{U}, \boldsymbol{\xi}, \xi') = \hat{\mu}_{\boldsymbol{\xi}} \left(\xi' : \rho_g(\xi' \odot \mathbf{U}) < \rho_g(\xi_k \odot \mathbf{U}) \right),$$

$$\hat{T}_v(\mathbf{U}, \boldsymbol{\xi}) = \hat{\mu}_{\boldsymbol{\xi}} \left(\xi' : \rho_v(\mathbf{U}) \leq \rho_v(\xi' \odot \mathbf{U}) \right),$$

and

$$\hat{T}_v^{\theta_G, \varepsilon_G}(\mathbf{U}, \boldsymbol{\xi}) = \frac{1}{\theta_G} \hat{\mu}_{\boldsymbol{\xi}} \left(\xi' : \rho_v(\mathbf{U}) \leq \rho_v(\xi' \odot \mathbf{U}); \check{T}_g^-(\mathbf{U}, \boldsymbol{\xi}, \xi') \leq \theta_G + \varepsilon_G \right).$$

We denote by $G_\beta(x, a, b)$ the c.d.f. of a beta distribution with parameters a and b evaluated at $x \in [0, 1]$, i.e.,

$$G_\beta(x, a, b) = \frac{1}{\beta(a, b)} \int_0^x t^{a-1}(1-t)^{b-1} dt$$

with $\beta(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$. We recall that if X is binomial with parameter n and p ($X \sim \mathrm{Bin}(n,p)$) then, for an integer $t \in \{0,\ldots,n\}$

$$P(X \leq t) = G_\beta(1-p, n-t, t+1).$$

We can now define

$$\hat{A} = \left\{ v : \hat{T}_{g(v)}(\mathbf{U}, \boldsymbol{\xi}) \leq \theta_G - \varepsilon_G \text{ and } \hat{T}_v^{\theta_G,\varepsilon_G}(\mathbf{U}, \boldsymbol{\xi}) \leq \theta_V \text{ and } \hat{T}_v(\mathbf{U}, \boldsymbol{\xi}) \leq \theta'_V \right\}$$

and state:

Theorem 2.2.2. *For $v \in V_0$,*

$$\mathbb{P}\left(v \in \hat{A}\right) \leq \frac{\lfloor K\theta'_V \rfloor + 1}{K+1}. \tag{2.11}$$

and, for $v \in V_{00}$ and $g = g(v)$,

$$\mathbb{P}\left(v \in \hat{A}\right) \leq c_K(\theta_G, \varepsilon_G) + \theta_G\theta_V \tag{2.12}$$

where

$$c_K(\theta_G, \varepsilon_G) = \frac{\lfloor K(\theta_G - \varepsilon_G) \rfloor + 1}{K+1} G_\beta(1-\theta_G, K - \lfloor K(\theta_G - \varepsilon_G) \rfloor, \lfloor K(\theta_G - \varepsilon_G) \rfloor + 2) \tag{2.13}$$

$$- \theta_G G_\beta(1-\theta_G, K - \lfloor K(\theta_G - \varepsilon_G) \rfloor, \lfloor K(\theta_G - \varepsilon_G) \rfloor + 1)$$

$$+ \theta_G G_\beta\big(\theta_G, \lfloor K(\theta_G + \varepsilon_G) \rfloor, K - \lfloor K(\theta_G + \varepsilon_G) \rfloor\big).$$

Here, $\lfloor x \rfloor$ denotes the integer part of x.

Corollary 2.2.2. *The FWER using the finite resampling scores is controlled by :*

$$\text{FWER} \leq |V| c_K(\theta_G, \varepsilon_G) + |V| \theta_G \theta_V + J \frac{\lfloor K\theta'_V \rfloor + 1}{K+1}.$$

Neglecting the rounding error in the last term (letting $(\lfloor K\theta'_V \rfloor + 1)(K+1) \simeq \theta'_V$), this theorem therefore adds the finite-sample correction $c_K(\theta_G, \varepsilon_G)$ to the asymptotic upper bound (theorem 2.2.1). Figure 2.1 plots the level curves of the logarithm of this correction as a function of K and ε_G, fixing θ_G to values that are used in our simulations.

The proof of Theorem 2.2.2 is based on the introduction of randomized finite sampling scores, allowing us to use lemma 2.2.1 at multiple occurrences. These randomized scores will be less conservative (but significantly more complex) than the scores that were introduced before Theorem 2.2.2, which will therefore be obtained

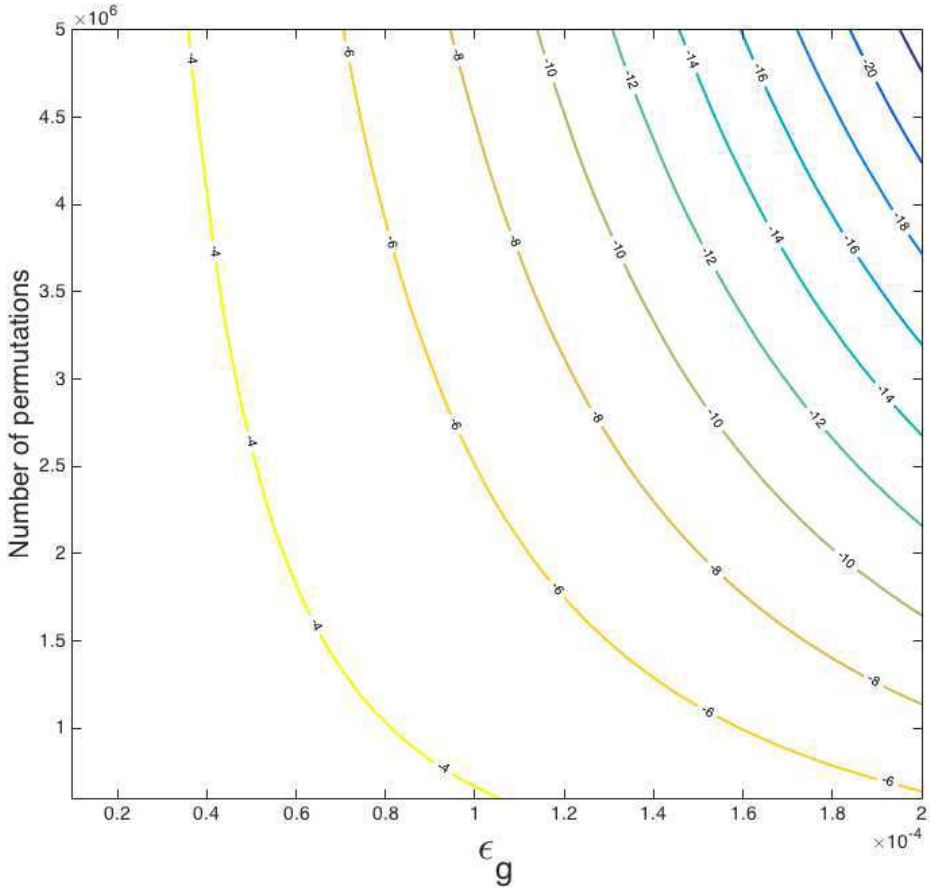

Figure 2.1: Level curves of the logarithm of $c_K(\theta_G, \varepsilon_G)$, with $\theta_G = 2.1 \times 10^{-3}$

as a corollary of the present proof.

Let us first recall our notation and introduce some new one. For a positive integer K, we let μ be the uniform probability on $\mathfrak{S}$ and let $\boldsymbol{\xi} = (\xi_1, ..., \xi_K) \in \mathfrak{S}^K$ where $\xi_1, \ldots, \xi_K$ are independent and have distribution μ. Also, we define $\mathbf{Z} = (Z_1, ..., Z_K)$ and $\tilde{\mathbf{Z}} = (\tilde{Z}_1, ..., \tilde{Z}_K)$ where $Z_1, \ldots, Z_K$ and $\tilde{Z}_1, \ldots, \tilde{Z}_K$ are independent random variables uniformly distributed on $[0, 1]$, that are also independent of $\mathbf{U}$ and $\boldsymbol{\xi}$. We will also need two additional independent uniformly distributed variables, Z and $\tilde{Z}$, also independent of all other variables. All these variables are assumed to be defined on a probability space $(\Omega, \mathbb{P})$. We also introduce the empirical measures

$$\hat{\mu}_{\boldsymbol{\xi}} = \frac{1}{K}\sum_{k=1}^{K} \delta_{\xi_k}, \quad \hat{\mu}_{\boldsymbol{\xi},\mathbf{Z}} = \frac{1}{K}\sum_{k=1}^{K} \delta_{\xi_k}\delta_{Z_k}, \text{ and } \hat{\mu}_{\boldsymbol{\xi},\tilde{\mathbf{Z}}} = \frac{1}{K}\sum_{k=1}^{K} \delta_{\xi_k}\delta_{\tilde{Z}_k}.$$

With these notations, we let:

$$\begin{aligned}
&\hat{T}_{g,r}(\mathbf{U}, \boldsymbol{\xi}, \mathbf{Z}, Z) \\
&= \hat{\mu}_{\boldsymbol{\xi}} \left(\xi' : \rho_g(\mathbf{U}) < \rho_g(\xi' \odot \mathbf{U})\right) + \hat{\mu}_{\boldsymbol{\xi},\mathbf{Z}} \left((\xi', z') : \rho_g(\mathbf{U}) = \rho_g(\xi' \odot \mathbf{U}); z' \leq Z\right) \\
&= \frac{1}{K}\sum_{k=1}^{K} \left(\mathbf{1}_{\rho_g(\mathbf{U}) < \rho_g(\xi_k \odot \mathbf{U})} + \mathbf{1}_{\rho_g(\mathbf{U}) = \rho_g(\xi_k \odot \mathbf{U})} \mathbf{1}_{Z_k \leq Z}\right)
\end{aligned}$$

$$\begin{aligned}
\check{T}_{g,r}(\mathbf{U}, \boldsymbol{\xi}, \mathbf{Z}_K, \xi', Z) = \hat{T}_g(\xi' \odot \mathbf{U}, \boldsymbol{\xi} \circ \xi'^{-1}, \mathbf{Z}, Z) \\
= \frac{1}{K}\sum_{k=1}^{K} \left(\mathbf{1}_{\rho_g(\xi' \odot \mathbf{U}) < \rho_g(\xi_k \odot \mathbf{U})} + \mathbf{1}_{\rho_g(\xi' \odot \mathbf{U}) = \rho_g(\xi_k \odot \mathbf{U})} \mathbf{1}_{Z_k \leq Z}\right),
\end{aligned}$$

$$\hat{T}_{v,r}(\mathbf{U}, \boldsymbol{\xi}, \mathbf{Z}, Z) = \hat{\mu}_{\boldsymbol{\xi}} \left(\xi' : \rho_v(\mathbf{U}) < \rho_v(\xi' \odot \mathbf{U})\right) + \hat{\mu}_{\boldsymbol{\xi},\mathbf{Z}} \left((\xi', z') : \rho_v(\mathbf{U}) = \rho_v(\xi' \odot \mathbf{U}); z' \leq Z\right)$$

and

$$\begin{aligned}
&\hat{T}^{\theta_G,\varepsilon_G}_{v,r}(\mathbf{U}, \boldsymbol{\xi}, \mathbf{Z}, \tilde{\mathbf{Z}}, Z, \tilde{Z}) \\
&= \frac{1}{\theta_G} \hat{\mu}_{\boldsymbol{\xi}} \left(\xi' : \rho_v(\mathbf{U}) < \rho_v(\xi' \odot \mathbf{U}); \check{T}_g(\mathbf{U}, \boldsymbol{\xi}, \mathbf{Z}, \xi', Z) \leq \theta_G + \varepsilon_G\right) \\
&+ \frac{1}{\theta_G} \hat{\mu}_{\boldsymbol{\xi},\tilde{\mathbf{z}}} \left((\xi', z') : \rho_v(\mathbf{U}) = \rho_v(\xi' \odot \mathbf{U}); z' \leq \tilde{Z}; \check{T}_g(\mathbf{U}, \boldsymbol{\xi}, \mathbf{Z}, \xi', Z) \leq \theta_G + \varepsilon_G\right)
\end{aligned}$$

We can now define:

$$\begin{aligned}
\hat{A}_r = \Big\{ v : \hat{T}_{g(v),r}(\mathbf{U}, \boldsymbol{\xi}, \mathbf{Z}, Z) \leq \theta_G - \varepsilon_G \text{ and } \hat{T}^{\theta_G,\varepsilon_G}_{v,r}(\mathbf{U}, \boldsymbol{\xi}, \mathbf{Z}, Z) \leq \theta_V \\
\text{and } \hat{T}_{v,r}(\mathbf{U}, \boldsymbol{\xi}, \mathbf{Z}, \tilde{\mathbf{Z}}, Z, \tilde{Z}) \leq \theta'_V \Big\}
\end{aligned}$$

It is easy to see that $\hat{A} \subset \hat{A}_r$. Therefore, the following result implies Theorem 2.2.2.

Theorem 2.2.3. *For $v \in V_0$,*

$$\mathbb{P}\left(v \in \hat{A}_r\right) \leq \frac{\lfloor K\theta'_V \rfloor + 1}{K+1}. \tag{2.14}$$

and, for $v \in V_{00}$ *and* $g = g(v)$,

$$\mathbb{P}\left(v \in \hat{A}_r\right) \leq \frac{\lfloor K(\theta_G - \varepsilon_G)\rfloor + 1}{K+1} G_\beta(1-\theta_G, K - \lfloor K(\theta_G-\varepsilon_G)\rfloor, \lfloor K(\theta_G-\varepsilon_G)\rfloor + 2)$$
$$- \theta_G G_\beta(1-\theta_G, K - \lfloor K(\theta_G-\varepsilon_G)\rfloor, \lfloor K(\theta_G-\varepsilon_G)\rfloor + 1)$$
$$+ \theta_G G_\beta\big(\theta_G, \lfloor K(\theta_G+\varepsilon_G)\rfloor, K - \lfloor K(\theta_G+\varepsilon_G)\rfloor + 1\big) + \theta_G\theta_V. \quad (2.15)$$

Proof of Theorem 2.2.3.

Step 1. We start with (2.14) which is simpler and standard. Let $v \in V_0$. Conditionally to $\mathbf{U}$ and Z, $K\hat{T}_{v,r}(\mathbf{U}, \boldsymbol{\xi}, \mathbf{Z}, Z)$ follows a binomial distribution $\mathrm{Bin}(K, T_v(\mathbf{U}, Z))$ (with T_v defined by equation (2.4)), so that

$$\mathbb{P}(v \in \hat{A}_r) \leq \mathbb{P}(\hat{T}_{v,r}(\mathbf{U}, \boldsymbol{\xi}, \mathbf{Z}, Z) \leq \theta'_V) = \mathbb{E}\left(G_\beta(1 - T_v(\mathbf{U}, Z), K - \lfloor K\theta'_V\rfloor, \lfloor K\theta'_V\rfloor + 1)\right).$$

Theorem 2.2.1 states that $T_v(\mathbf{U}, Z)$ follows a uniform distribution on $[0, 1]$. Therefore,

$$\mathbb{P}(v \in \hat{A}_r) \leq \int_0^1 G_\beta(t, K - \lfloor K\theta'_V\rfloor, \lfloor K\theta'_V\rfloor + 1)dt = \frac{\lfloor K\theta'_V\rfloor + 1}{K+1}.$$

Step 2. We now consider (2.15) and take $v \in V_{00}$, $g = g(v)$. We first prove that:

$$\begin{aligned}
&\mathbb{P}\left(\hat{T}_{g,r}(\mathbf{U},\boldsymbol{\xi},\mathbf{Z},Z) \le \theta_G - \varepsilon_G;\, \hat{T}_{v,r}^{\theta_G,\varepsilon_G}(\mathbf{U},\boldsymbol{\xi},\mathbf{Z},\tilde{\mathbf{Z}},Z,\tilde{Z}) \le \theta_V\right) \qquad (2.16)\\
&\quad \le \mathbb{P}\left(T_g(\mathbf{U},Z) \le \theta_G;\, \hat{T}_{v,r}^{\theta_G,\varepsilon_G}(\mathbf{U},\boldsymbol{\xi},\mathbf{Z},\tilde{\mathbf{Z}},Z,\tilde{Z}) \le \theta_V\right)\\
&\quad + \frac{\lfloor K(\theta_G-\varepsilon_G)\rfloor + 1}{K+1} G_\beta(1-\theta_G, K - \lfloor K(\theta_G-\varepsilon_G)\rfloor, \lfloor K(\theta_G-\varepsilon_G)\rfloor + 2)\\
&\quad - \theta_G G_\beta(1-\theta_G, K - \lfloor K(\theta_G-\varepsilon_G)\rfloor, \lfloor K(\theta_G-\varepsilon_G)\rfloor + 1)
\end{aligned}$$

Notice that:

$$\begin{aligned}
&\mathbb{P}\left(\hat{T}_{g,r}(\mathbf{U},\boldsymbol{\xi},\mathbf{Z},Z) \le \theta_G - \varepsilon_G;\, \hat{T}_{v,r}^{\theta_G,\varepsilon_G}(\mathbf{U},\boldsymbol{\xi},\mathbf{Z},\tilde{\mathbf{Z}},Z,\tilde{Z}) \le \theta_V\right) -\\
&\qquad \mathbb{P}\left(T_g(\mathbf{U},Z) \le \theta_G;\, \hat{T}_{v,r}^{\theta_G,\varepsilon_G}(\mathbf{U},\boldsymbol{\xi},\mathbf{Z},\tilde{\mathbf{Z}},Z,\tilde{Z}) \le \theta_V\right)\\
&\qquad\qquad \le \mathbb{P}\left(\hat{T}_{g,r}(\mathbf{U},\boldsymbol{\xi},\mathbf{Z},Z) \le \theta_G - \varepsilon_G;\, T_g(\mathbf{U},Z) \ge \theta_G\right)
\end{aligned}$$

Now, write:

$$\begin{aligned}
&\mathbb{P}\left(\hat{T}_{g,r}(\mathbf{U},\boldsymbol{\xi},\mathbf{Z},Z) \le \theta_G - \varepsilon_G;\, T_g(\mathbf{U},Z) \ge \theta_G\right)\\
&\qquad = \mathbb{E}\left(\mathbb{P}\left(\hat{T}_{g,r}(\mathbf{U},\boldsymbol{\xi},\mathbf{Z},Z) \le \theta_G - \varepsilon_G;\, T_g(\mathbf{U},Z) \ge \theta_G \,|\, \mathbf{U},Z\right)\right).
\end{aligned}$$

By the same remark used to prove (2.14), we have that:

$$\mathbb{P}\left(\hat{T}_{g,r}(\mathbf{U},\boldsymbol{\xi},\mathbf{Z},Z) \leq \theta_G - \varepsilon_G;\, T_g(\mathbf{U},Z) \geq \theta_G \,|\, \mathbf{U},Z\right)$$
$$= \mathbf{1}_{T_g(\mathbf{U},Z)\geq\theta_G} G_\beta(1 - T_g(\mathbf{U},Z), K - \lfloor K(\theta_G - \varepsilon_G)\rfloor, \lfloor K(\theta_G - \varepsilon_G)\rfloor + 1).$$

Taking the expectation over $\mathbf{U}$ and Z, and using the fact that $1 - T_g(\mathbf{U},Z)$ is uniformly distributed over $[0,1]$ under the "double null" hypothesis:

$$\begin{aligned}
&\mathbb{P}\left(\hat{T}_{g,r}(\mathbf{U},\boldsymbol{\xi},\mathbf{Z}_K,Z) \leq \theta_G - \varepsilon_G;\, T_g(\mathbf{U},Z) \geq \theta_G\right) \\
&= \int_{\theta_G}^{1} G_\beta(1-t, K - \lfloor K(\theta_G - \varepsilon_G)\rfloor, \lfloor K(\theta_G - \varepsilon_G)\rfloor + 1)dt \\
&= \frac{\lfloor K(\theta_G - \varepsilon_G)\rfloor + 1}{K+1} G_\beta(1-\theta_G, K - \lfloor K(\theta_G - \varepsilon_G)\rfloor, \lfloor K(\theta_G - \varepsilon_G)\rfloor + 2) \\
&- \theta_G G_\beta(1-\theta_G, K - \lfloor K(\theta_G - \varepsilon_G)\rfloor, \lfloor K(\theta_G - \varepsilon_G)\rfloor + 1)
\end{aligned}$$

Step 3. Let

$$\begin{aligned}\tilde{T}_{v,r}^{\theta_G}(\mathbf{U},\boldsymbol{\xi},\mathbf{Z},\tilde{\mathbf{Z}},Z,\tilde{Z}) &= \frac{1}{K\theta_G}\sum_{k=1}^{K}\mathbf{1}_{\rho_v(\mathbf{U})<\rho_v(\xi_k\odot\mathbf{U})}\mathbf{1}_{T_g(\xi_k\odot\mathbf{U},Z)\leq\theta_G}\\ &\quad+\frac{1}{K\theta_G}\sum_{k=1}^{K}\mathbf{1}_{\rho_v(\mathbf{U})=\rho_v(\xi_k\odot\mathbf{U})}\mathbf{1}_{\tilde{Z}_k\leq\tilde{Z}}\mathbf{1}_{T_g(\xi_k\odot\mathbf{U},Z)\leq\theta_G}\\ &= \frac{1}{K\theta_G}\hat{\mu}_{\boldsymbol{\xi}}\left(\xi':\rho_v(\xi'\odot\mathbf{U})>\rho_v(\mathbf{U});\,T_g(\xi'\odot\mathbf{U},Z)\leq\theta_G\right)\\ &\quad+\frac{1}{K\theta_G}\hat{\mu}_{\boldsymbol{\xi},\tilde{\mathbf{Z}}}\left((\xi',z'):\rho_v(\xi'\odot\mathbf{U})=\rho_v(\mathbf{U});z'\leq\tilde{Z}\,T_g(\xi'\odot\mathbf{U},Z)\leq\theta_G\right)\end{aligned}$$

We now prove that

$$\begin{aligned}&\mathbb{P}\left(T_g(\mathbf{U},Z)\leq\theta_G;\,\hat{T}_{v,r}^{\theta_G,\varepsilon_G}(\mathbf{U},\boldsymbol{\xi},\mathbf{Z},\tilde{\mathbf{Z}},Z,\tilde{Z})\leq\theta_V\right)\\ &\qquad\leq\mathbb{P}\left(T_g(\mathbf{U},Z)\leq\theta_G;\,\tilde{T}_{v,r}^{\theta_G}(\mathbf{U},\boldsymbol{\xi},\mathbf{Z},\tilde{\mathbf{Z}},Z,\tilde{Z})\leq\theta_V\right)\\ &\qquad\qquad+\theta_G G_\beta(\theta_G,\lfloor K(\theta_G+\varepsilon_G)\rfloor+1,K+1).\end{aligned}$$

Notice that:

$$\begin{aligned}&\mathbb{P}\left(T_g(\mathbf{U},Z)\leq\theta_G;\,\hat{T}_{v,r}^{\theta_G,\varepsilon_G}(\mathbf{U},\boldsymbol{\xi},\mathbf{Z},\tilde{\mathbf{Z}},Z,\tilde{Z})\leq\theta_V\right)-\mathbb{P}\left(T_g(\mathbf{U},Z)\leq\theta_G;\,\tilde{T}_{v,r}^{\theta_G}(\mathbf{U},\boldsymbol{\xi},\mathbf{Z},\tilde{\mathbf{Z}},Z,\tilde{Z})\leq\theta_V\right)\\ &\qquad\leq\mathbb{P}\left(T_g(\mathbf{U},Z)\leq\theta_G;\,\hat{T}_{v,r}^{\theta_G,\varepsilon_G}(\mathbf{U},\boldsymbol{\xi},\mathbf{Z},\tilde{\mathbf{Z}},Z,\tilde{Z})\leq\tilde{T}_{v,r}^{\theta_G}(\mathbf{U},\boldsymbol{\xi},\mathbf{Z},\tilde{\mathbf{Z}},Z,\tilde{Z})\right)\quad(2.17)\end{aligned}$$

Conditioning on $\mathbf{U}$ and Z taking the expected value:

$$\mathbb{P}\left(T_g(\mathbf{U},Z)\leq\theta_G;\,\hat{T}^{\theta_G,\varepsilon_G}_{v,r}(\mathbf{U},\boldsymbol{\xi},\mathbf{Z},\tilde{\mathbf{Z}},Z,\tilde{Z})\leq\tilde{T}^{\theta_G}_{v,r}(\mathbf{U},\boldsymbol{\xi},\mathbf{Z},\tilde{\mathbf{Z}},Z,\tilde{Z})\right)$$
$$=\mathbb{E}\left(\mathbf{1}_{T_g(\mathbf{U},Z)\leq\theta_G}\mathbb{P}\left(\hat{T}^{\theta_G,\varepsilon_G}_{v,r}(\mathbf{U},\boldsymbol{\xi},\mathbf{Z},\tilde{\mathbf{Z}},Z,\tilde{Z})\leq\tilde{T}^{\theta_G}_{v,r}(\mathbf{U},\boldsymbol{\xi},\mathbf{Z},\tilde{\mathbf{Z}},Z,\tilde{Z})\,|\,\mathbf{U},Z\right)\right)$$

We rewrite the last expectation as:

$$\mathbb{E}\Bigg(\mathbf{1}_{T_g(\mathbf{U},Z)\leq\theta_G}\mathbb{P}\Bigg(\sum_{k=1}^{K}(\mathbf{1}_{\rho_v(\mathbf{U})<\rho_v(\xi_k\odot\mathbf{U})}+\mathbf{1}_{\rho_v(\mathbf{U})=\rho_v(\xi_k\odot\mathbf{U})}\mathbf{1}_{\tilde{Z}_k\leq\tilde{Z}})$$
$$\times(\mathbf{1}_{\check{T}_{g,r}(\mathbf{U},\boldsymbol{\xi},\xi_k,\mathbf{Z},Z)\leq\theta_G+\varepsilon_G}-\mathbf{1}_{T_g(\xi_k\odot\mathbf{U},Z)\leq\theta_G})\leq 0\,\Big|\,\mathbf{U},Z\Bigg)\Bigg).$$

In order that

$$\sum_{k=1}^{K}(\mathbf{1}_{\rho_v(\mathbf{U})<\rho_v(\xi_k\odot\mathbf{U})}+\mathbf{1}_{\rho_v(\mathbf{U})=\rho_v(\xi_k\odot\mathbf{U})}\mathbf{1}_{\tilde{Z}_k\leq\tilde{Z}})$$
$$\times(\mathbf{1}_{\check{T}_{g,r}(\mathbf{U},\boldsymbol{\xi},\xi_k,\mathbf{Z},Z)\leq\theta_G+\varepsilon_G}-\mathbf{1}_{T_g(\xi_k\odot\mathbf{U},Z)\leq\theta_G})\leq 0,\quad(2.18)$$

there must exist $k_0\in\{1,2,...,K\}$ such that $\check{T}_{g,r}(\mathbf{U},\boldsymbol{\xi},\xi_k,\mathbf{Z},Z)>\theta_G+\varepsilon_G$ and $T_g(\xi_k\odot\mathbf{U},Z)\leq\theta_G$. Letting $M_g(\boldsymbol{\xi},\mathbf{U},\mathbf{Z})$ denote the number of indexes k such that $T_g(\xi_k\odot\mathbf{U},Z_k)\leq\theta_G$, the existence of such a k_0 implies that $M_g(\boldsymbol{\xi},\mathbf{U},\mathbf{Z})>K(\theta_G+\varepsilon_G)$. Indeed, we first notice that, for a every $0\leq j,k\leq K$:

$$\mathbf{1}_{\rho_g(\xi_k\odot\mathbf{U})<\rho_g(\xi_j\odot\mathbf{U})}+\mathbf{1}_{\rho_g(\xi_k\odot\mathbf{U})=\rho_g(\xi_j\odot\mathbf{U})}\mathbf{1}_{Z_j\leq Z}\leq\mathbf{1}_{T_g(\xi_j\odot\mathbf{U},Z_j)\leq T_g(\xi_k\odot\mathbf{U},Z)}.$$

This statement is obvious when $\rho_g(\xi_k \odot \mathbf{U}) = \rho_g(\xi_j \odot \mathbf{U})$ and $Z_j \leq Z$ and can be checked by proving that $T_g(\xi_j \odot \mathbf{U}, 1) \leq T_g(\xi_k \odot \mathbf{U}, 0)$ when $\rho_g(\xi_k \odot \mathbf{U}) < \rho_g(\xi_j \odot \mathbf{U})$. Therefore, if k_0 exists, we must have

$$\begin{aligned}\check{T}_{g,r}(\mathbf{U}, \boldsymbol{\xi}, \xi_{k_0}, \mathbf{Z}, Z) &\leq \frac{1}{K}\sum_{j=1}^{K} \mathbf{1}_{T_g(\xi_j \odot \mathbf{U}, Z_j) \leq T_g(\xi_{k_0} \odot \mathbf{U}, Z)} \\ &\leq \frac{1}{K}\sum_{j=1}^{K} \mathbf{1}_{T_g(\xi_j \odot \mathbf{U}, Z_j) \leq \theta_G}.\end{aligned}$$

Because $\mathbb{E}\left(\mu(T_g(\xi \odot \mathbf{U}, Z) \leq \theta_G)|\mathbf{U}\right) = \theta_G$, we can bound the probability of (2.18) conditional to $\mathbf{U}$ and Z by the probability that a binomial $\mathrm{Bin}(K, \theta_G)$ is larger than $\lceil K(\theta_G + \varepsilon_G)\rceil$, which is given by

$$G_\beta(\theta_G, \lfloor K(\theta_G + \varepsilon_G)\rfloor + 1, K - \lfloor K(\theta_G + \varepsilon_G)\rfloor).$$

We therefore have

$$\begin{aligned}\mathbb{P}\left(\hat{T}_{v,r}^{\theta_G, \varepsilon_G}(\mathbf{U}, \boldsymbol{\xi}, \mathbf{Z}, \tilde{\mathbf{Z}}, Z, \tilde{Z}) \leq \tilde{T}_{v,r}^{\theta_G}(\mathbf{U}, \boldsymbol{\xi}, \mathbf{Z}, \tilde{\mathbf{Z}}, Z, \tilde{Z}) \,|\, \mathbf{U}, Z\right) \\ \leq G_\beta(\theta_G, \lfloor K(\theta_G + \varepsilon_G)\rfloor + 1, K - \lfloor K(\theta_G + \varepsilon_G)\rfloor).\end{aligned}$$

Now, finally notice that

$$\mathbb{P}\left(T_g(\mathbf{U}, Z) \leq \theta_G\right) = \theta_G,$$

and we have proved (2.17).

Step 4. We finally show that

$$\mathbb{P}\left(T_g(\mathbf{U},Z)\leq\theta_G;\,\tilde{T}_{v,r}^{\theta_G}(\mathbf{U},\boldsymbol{\xi},\mathbf{Z},\tilde{\mathbf{Z}},Z,\tilde{Z})\leq\theta_V\right)\leq\theta_V\theta_G \tag{2.19}$$

For this, we note that

$$\begin{aligned}K\theta_G\tilde{T}_{v,r}^{\theta_G}(\mathbf{U},\boldsymbol{\xi},\mathbf{Z},\tilde{\mathbf{Z}},Z,\tilde{Z}) &= \sum_{k=1}^{K}\mathbf{1}_{\rho_v(\mathbf{U})<\rho_v(\xi_k\odot\mathbf{U})}\mathbf{1}_{T_g(\xi_k\odot\mathbf{U},Z)\leq\theta_G}\\ &\quad+\sum_{k=1}^{K}\mathbf{1}_{\rho_v(\mathbf{U})=\rho_v(\xi_k\odot\mathbf{U})}\mathbf{1}_{\tilde{Z}_k\leq\tilde{Z}}\mathbf{1}_{T_g(\xi_k\odot\mathbf{U},Z)\leq\theta_G}.\end{aligned}$$

Conditionally to $\mathbf{U}$, Z and $\tilde{Z}$, this variable follows a binomial distribution with probability of success $N_g^{\theta_G}(\mathbf{U},Z)T_v^{\theta_G}(\mathbf{U},Z,\tilde{Z})$. Therefore:

$$\begin{aligned}&\mathbb{P}\left(T_g(\mathbf{U},Z)\leq\theta_G;\tilde{T}_{v,r}^{\theta_G}(\mathbf{U},\boldsymbol{\xi},\mathbf{Z},\tilde{\mathbf{Z}},Z,\tilde{Z})\leq\theta_V\right)\\ &=\mathbb{E}\left(\mathbf{1}_{T_g(\mathbf{U},Z)\leq\theta_G}G_\beta(1-N_g^{\theta_G}(\mathbf{U},Z)T_v^{\theta_G}(\mathbf{U},Z,\tilde{Z}),K-\lfloor K\theta_G\theta_V\rfloor,\lfloor K\theta_G\theta_V\rfloor+1)\right)\end{aligned}$$

We now use the fact that The distribution of $\mathbf{U}$ is invariant under the action of the group $\mathfrak{S}$ and that $N_g^{\theta_G}(\xi\odot\mathbf{U},Z)=N_g^{\theta_G}(\mathbf{U},Z)$ for all $\xi\in\mathfrak{S}$ to write, introducing a

new random variable $\bar{\xi}$ independent from the others in the expectation

$$\begin{aligned}
&\mathbb{P}\left(T_g(\mathbf{U},Z)\leq\theta_G;\tilde{T}^{\theta_G}_{v,r}(\mathbf{U},\boldsymbol{\xi},\mathbf{Z},\tilde{\mathbf{Z}},Z,\tilde{Z})\leq\theta_V\right)\\
&=\mathbb{E}\left(\mathbf{1}_{T_g(\bar{\xi}\odot\mathbf{U},Z)\leq\theta_G}G_\beta(1-N^{\theta_G}_g(\mathbf{U},Z)T^{\theta_G}_v(\bar{\xi}\odot\mathbf{U},Z,\tilde{Z}),K-\lfloor K\theta_G\theta_V\rfloor,\lfloor K\theta_G\theta_V\rfloor+1)\right)
\end{aligned}$$

Now, using lemma 2.2.1, we notice that, given $\mathbf{U}$ and Z, the random variable $T^{\theta_G}_v(\xi\odot\mathbf{U},Z,\tilde{Z})$ is, conditionally to $T_g(\xi\odot\mathbf{U},Z)\leq\theta_G$, uniformly distributed over $[0,1]$. Recall also that $N^{\theta_G}_g(\mathbf{U},Z)$ is, by definition, equal to $\mathbb{P}(T_g(\bar{\xi}\odot\mathbf{U},Z)\leq\theta_G|U,Z)$. From this, it follows that

$$\begin{aligned}
&\mathbb{P}\left(T_g(\mathbf{U},Z)\leq\theta_G;\tilde{T}^{\theta_G}_{v,r}(\mathbf{U},\boldsymbol{\xi},\mathbf{Z},\tilde{\mathbf{Z}},Z,\tilde{Z})\leq\theta_V\right)\\
&=\mathbb{E}\left(N^{\theta_G}_g(\mathbf{U},Z)\int_0^1 G_\beta(1-N^{\theta_G}_g(\mathbf{U},Z)t,Z,\tilde{Z}),K-\lfloor K\theta_G\theta_V\rfloor,\lfloor K\theta_G\theta_V\rfloor+1)dt\right)\\
&=\mathbb{E}\left(N^{\theta_G}_g(\mathbf{U},Z)\int_0^1\int_0^{1-N^{\theta_G}_g(\mathbf{U},Z)t}s^{K-\lfloor K\theta_G\theta_V\rfloor-1}(1-s)^{\lfloor K\theta_G\theta_V\rfloor}dsdt\right)\\
&=\mathbb{E}\left(N^{\theta_G}_g(\mathbf{U},Z)\int_0^1 s^{K-\lfloor K\theta_G\theta_V\rfloor-1}(1-s)^{\lfloor K\theta_G\theta_V\rfloor}\int_0^{\min\left(\frac{1-s}{N^{\theta_G}_g(\mathbf{U},Z)},1\right)}dtds\right)\\
&\leq\mathbb{E}\left(N^{\theta_G}_g(\mathbf{U},Z)\int_0^1 s^{K-\lfloor K\theta_G\theta_V\rfloor-1}(1-s)^{\lfloor K\theta_G\theta_V\rfloor}\int_0^{\frac{1-s}{N^{\theta_G}_g(\mathbf{U},Z)}}dtds\right)\\
&=\mathbb{E}\left(\int_0^1 s^{K-\lfloor K\theta_G\theta_V\rfloor-1}(1-s)^{\lfloor K\theta_G\theta_V\rfloor+1}ds\right)\\
&=\frac{\lfloor K\theta_G\theta_V\rfloor+1}{K+1}.
\end{aligned}$$

Hence we proved (2.19), which finishes the proof of lemma 2.2.3 and of theorem 2.2.2.

□

2.3 Estimating the number of indices inside active cells

We now focus on the issue of estimating from observed data the number J of indices belonging to active cells.

2.3.1 Asymptotic resampling scores

Recall that:

$$J = \sum_{g \in G_0^c} |g|,$$

where G_0 is the set of inactive cells. Our estimation will be made based on cell statistics $(T_g(\mathbf{U}), g \in G)$ under the following assumption. We will assume that T_g takes small values when g is active, so that, for a suitable non conservative threshold t_0, we have $P(T_g(\mathbf{U}) \leq t_0) \simeq 1$. To simplify the argument, we will actually make the approximation that:

[A] There exists $t_0 \in (0, 1)$ such that $\mathbb{P}(T_g(\mathbf{U}) \leq t_0) = 1$ if $g \cap A \neq \emptyset$.

This assumption is sometimes called the zero assumption (see for example[22]) because it is assuming that the probability for each of the scores at the cell level, when the cells

are non active, to be in a certain region is zero. The zero assumption is usually made at the indices level and is common under the Bayesian multiple testing framework.[22] Notice that if we denote by:

$$G_0(\mathbf{U}) = \{g \in G : T_g(\mathbf{U}) > t_0\},$$

then assumption [A] implies that $G_0(\mathbf{U}) \subset G_0$. This in turn implies that:

$$N_0(\mathbf{U}) := \sum_{g \in G_0(\mathbf{U})} |g| \leq \sum_{g \in G_0} |g| = |V| - J.$$

Assumption [A] therefore implies an estimator for a lower bound for $|V| - J$ and therefore an upper bound for J, with holds with probability one. However, since the choice of t_0 will not be conservative (typically greater than 0.25), this upper bound will not be sharp enough to be able to take advantage of the clustering assumption. The purpose of this part is to use the set $G_0(\mathbf{U})$ to derive a less obvious and sharper upper bound of J. We start by defining the statistics that will be used to derive the estimator. We will as usual denote our group of transformations by $\mathfrak{S}$, the elements of the group by ξ and the group action by $\odot$. We furthermore define:

- $\odot$ for each $\xi \in \mathfrak{S}$, $N_1(\mathbf{U}, \xi) = \sum_{g \in G_0(\mathbf{U})} \mathbf{1}_{T_g(\xi \odot \mathbf{U}) \leq t_0} |g|$,
- $\odot$ $q_1(\mathbf{U}, \epsilon) = \sup \{n \in \mathbb{N} : \mu(\xi : N_1(\mathbf{U}, \xi) \geq n) > 1 - \epsilon\}$.

$\odot$ Finally,

$$\hat{J}(\mathbf{U}, \epsilon) = |V| - (N_0(\mathbf{U}) + q_1(\mathbf{U}, \epsilon)) . \tag{2.20}$$

With this notation, we have the following main result.

Theorem 2.3.1. *Assuming [A], we have:*

$$\mathbb{P}\left(\hat{J}(\mathbf{U}, \epsilon) < J\right) \leq \epsilon. \tag{2.21}$$

Proof. First, let us remark that proving (2.21) is equivalent to proving:

$$\mathbb{P}\left(N_0(\mathbf{U}) + q_1(\mathbf{U}, \epsilon) \geq \sum_{g \in G_0} |g|\right) \leq \epsilon, \tag{2.22}$$

and that we can write

$$\mathbb{P}\left(N_0(\mathbf{U}) + q_1(\mathbf{U}, \epsilon) \geq \sum_{g \in G_0} |g|\right) = \mathbb{P}\left(q_1(\mathbf{U}, \epsilon) \geq \sum_{g \in G_0} \mathbf{1}_{T_g(\mathbf{U}) \leq t_0} |g|\right).$$

Now define:

$$\tilde{N}_1(\mathbf{U}) = \sum_{g \in G_0} \mathbf{1}_{T_g(\mathbf{U}) \leq t_0} |g|.$$

Since $G_0(\mathbf{U}) \subset G_0$, we have $\tilde{N}_1(\xi \odot \mathbf{U}) \geq N_1(\mathbf{U}, \xi)$ for every $\xi \in \mathfrak{S}$, and $\tilde{q}_1(\mathbf{U}, \epsilon)$ defined as:

$$\tilde{q}_1(\mathbf{U}, \epsilon) = \sup\left\{n \in \mathbb{N} : \mu\left(\xi : \tilde{N}_1(\xi \odot \mathbf{U}) \geq n\right) > 1 - \epsilon\right\}$$

satisfies $\tilde{q}_1(\mathbf{U}, \epsilon) \geq q_1(\mathbf{U}, \epsilon)$ so that

$$\mathbb{P}\left(q_1(\mathbf{U}, \epsilon) \geq \sum_{g \in G_0} \mathbf{1}_{T_g(\mathbf{U}) \leq t_0} |g| \right) \leq \mathbb{P}\left(\tilde{q}_1(\mathbf{U}, \epsilon) \geq \sum_{g \in G_0} \mathbf{1}_{T_g(\mathbf{U}) \leq t_0} |g| \right).$$

By noticing that $\tilde{q}_1(\mathbf{U}, \epsilon) = \tilde{q}_1(\xi \odot \mathbf{U}, \epsilon)$ for every $\xi \in \mathfrak{S}$ by definition of $\tilde{q}_1$, and that the distribution of $\sum_{g \in G_0} \mathbf{1}_{T_g(\mathbf{U}) \leq t_0} |g|$ is invariant under the action of any element $\xi \in \mathfrak{S}$, we have the following:

$$\mathbb{P}\left(\tilde{q}_1(\mathbf{U}, \epsilon) \geq \sum_{g \in G_0} \mathbf{1}_{T_g(\mathbf{U}) \leq t_0} |g| \right) = \mathbb{E}\left(\mu\left(\xi : \tilde{q}_1(\mathbf{U}, \epsilon) \geq \tilde{N}_1(\xi \odot \mathbf{U}) \right) \right).$$

But

$$\mu\left(\xi : \tilde{q}_1(\mathbf{U}, \epsilon) \geq \tilde{N}_1(\xi \odot \mathbf{U}) \right) \leq \epsilon$$

by definition of $\tilde{q}_1$ and (2.22) is proved. □

2.3.2 Finite resampling scores

We now discuss how the previous estimation of J can be modified when the uniform measure on $\mathfrak{S}$ is approximated by random sampling. Assuming two independent groups of i.i.d. samples of μ, $\boldsymbol{\xi} = (\xi_1, \xi_2, ..., \xi_K)$ and $\boldsymbol{\xi}' = (\xi'_1, \xi'_2, ...\xi'_{K'})$, and using the notation of section 2.2.3, we define:

$$\hat{T}_g(\mathbf{U}, \boldsymbol{\xi}) = \hat{\mu}_{\boldsymbol{\xi}}\left(\xi' : \rho_g(\mathbf{U}) \leq \rho_g(\xi' \odot \mathbf{U}) \right) = \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}_{\rho_g(\mathbf{U}) \leq \rho_g(\xi_k \odot \mathbf{U})}$$

and for every $i \in \{1, 2, ..., K'\}$:

$$\check{T}_g(\mathbf{U}, \boldsymbol{\xi}, \xi_i') = \hat{T}_g(\xi' \odot \mathbf{U}, \boldsymbol{\xi} \circ \xi_i'^{-1}) = \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}_{\rho_g(\xi_i' \odot \mathbf{U}) \leq \rho_g(\xi_k \odot \mathbf{U})}, .$$

We replace the assumption [A] of the previous section by the assumption $[\hat{A}]$:

$[\hat{A}]$ There exists $t_0 \in (0, 1)$ such that $\mathbb{P}(\hat{T}_g(\mathbf{U}, \boldsymbol{\xi}) \leq t_0) = 1$ if $g \cap A \neq \emptyset$.

Notice that is is possible to keep the previous assumption [A] and replace t_0 by $t_0 + \epsilon$ in $[\hat{A}]$ and the probability 1 by $1 - \exp(-2K\epsilon^2)$, using a Hoeffding bound. We now provide an upper bound $\hat{J}$ of the number of indices belonging to active cells using the finite resampling scores. We will use the following notation.

- $\odot$ Let $\hat{G}_0(\mathbf{U}, \boldsymbol{\xi}) = \{g \in G : \hat{T}_g(\mathbf{U}, \boldsymbol{\xi}) > t_0\}$. (Notice that assumption $[\hat{A}]$ implies that $\hat{G}_0(\mathbf{U}, \boldsymbol{\xi}) \subset G_0$).

- $\odot$ Let $\hat{N}_0(\mathbf{U}, \boldsymbol{\xi}) = \sum_{g \in G} |g| \mathbf{1}_{\hat{T}_g(\mathbf{U}, \boldsymbol{\xi}) > t_0}$ and $\hat{N}_1(\mathbf{U}, \boldsymbol{\xi}) = \sum_{g \in G_0} |g| \mathbf{1}_{\hat{T}_g(\mathbf{U}, \boldsymbol{\xi}) \leq t_0}$, so that

$$\hat{N}_0(\mathbf{U}, \boldsymbol{\xi}) + \hat{N}_1(\mathbf{U}, \boldsymbol{\xi}) = |V| - J.$$

- $\odot$ For each $i \in \{1, 2, ..., K'\}$, $\check{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi_i') = \sum_{g \in \hat{G}_0(\mathbf{U}, \boldsymbol{\xi})} |g| \mathbf{1}_{\check{T}_g(\mathbf{U}, \boldsymbol{\xi}, \xi_i') \leq t_0}$.

- $\odot$ The order statistics of the K' random variables

$$\check{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi_1'), \check{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi_2'), ..., \check{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi_{K'}'),$$

will be denoted by:

$$\check{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi'_{(1)}), \check{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi'_{(2)}), ..., \check{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi'_{(K')}).$$

($\check{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi'_{(1)})$ being the smallest statistic).

⊙ Finally, define

$$\hat{J}(\mathbf{U}, \boldsymbol{\xi}, \boldsymbol{\xi}', p) = |V| - \left(\hat{N}_0(\mathbf{U}, \boldsymbol{\xi}) + \check{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi'_{(p)})\right).$$

Notice that the computation of $\hat{J}(\mathbf{U}, \boldsymbol{\xi}, \boldsymbol{\xi}', p)$ requires the computation of just $K + K'$ scores.

We have the following result.

Theorem 2.3.2.

$$\mathbb{P}\left(\hat{J}(\mathbf{U}, \boldsymbol{\xi}, \boldsymbol{\xi}', p) < J\right) \leq \frac{p-1}{K'}.$$

Proof. We first notice that $\hat{J}(\mathbf{U}, \boldsymbol{\xi}, \boldsymbol{\xi}', p) = J + \hat{N}_1(\mathbf{U}, \boldsymbol{\xi}) - \check{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi'_{(p)})$, so that

$$\begin{aligned}\mathbb{P}\left(\hat{J}(\mathbf{U}, \boldsymbol{\xi}, \boldsymbol{\xi}', p) < J\right) &= \mathbb{P}\left(\check{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi'_{(p)}) > \hat{N}_1(\mathbf{U}, \boldsymbol{\xi})\right) \\ &\leq \mathbb{P}\left(\tilde{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi'_{(p)}) > \hat{N}_1(\mathbf{U}, \boldsymbol{\xi})\right)\end{aligned}$$

where for each $i \in \{1, 2, ..., K'\}$,

$$\tilde{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi_i^{'}) = \sum_{g \in G_0} |g| \mathbf{1}_{\check{T}_g(\mathbf{U}, \boldsymbol{\xi}, \xi_i^{'}) \leq t_0}$$

and $\tilde{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi_{(p)}^{'})$ is the corresponding pth order statistic. The last inequality holds because, for every j, $\tilde{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi_j) \geq \check{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi_j^{'})$, which implies

$$\tilde{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi_{(p)}^{'}) \geq \check{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi_{(p)}^{'})$$

We then have:

$$\boldsymbol{P}\left(\tilde{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi_{(p)}^{'}) > \hat{N}_1(\mathbf{u}, \boldsymbol{\xi})\right) = \boldsymbol{P}\left(|\{i : \tilde{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi_i^{'}) < \hat{N}_1(\mathbf{U}, \boldsymbol{\xi})\}| < p\right).$$

Notice that given $\boldsymbol{\xi}$ and $\mathbf{U}$, the variable $|\{i : \tilde{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi_i^{'}) < \hat{N}_1(\mathbf{u}, \boldsymbol{\xi})\}|$ follows a Binomial distribution with $K^{'}$ number of trials and a probability of success that is equal to:

$$\mu\left(\xi' : \tilde{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi^{'}) < \hat{N}_1(\mathbf{U}, \boldsymbol{\xi})\right).$$

Using the fact that $\check{T}_g(\mathbf{U}, \boldsymbol{\xi}, \xi^{'}) = \check{T}_g(\xi' \odot \mathbf{U}, \boldsymbol{\xi} \circ \xi^{'^{-1}})$, we have:

$$\tilde{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi^{'}) = \tilde{N}_1(\xi^{'} \odot \mathbf{U}, \boldsymbol{\xi} \circ \xi^{'^{-1}}),$$

and

$$\mu\left(\xi' : \tilde{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi') < \hat{N}_1(\mathbf{U}, \boldsymbol{\xi})\right) = \mu\left(\xi' : \tilde{N}_1(\xi' \odot \mathbf{U}, \boldsymbol{\xi} \circ \xi'^{-1}) < \hat{N}_1(\mathbf{U}, \boldsymbol{\xi})\right).$$

At this point, we notice that $\mathfrak{S}$ acts on $\mathcal{U} \times \mathfrak{S}^K$ via the group action:

$$\left(\xi', (\mathbf{U}, \boldsymbol{\xi})\right) \to (\xi' \odot \mathbf{U}, \boldsymbol{\xi} \circ \xi'^{-1}),$$

and this group action leaves invariant the joint distribution of $(\mathbf{U}, \boldsymbol{\xi})$. Therefore, the distribution of $\mu\left(\xi' : \tilde{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi') < \tilde{N}_1(\mathbf{U}, \boldsymbol{\xi})\right)$ is dominated by the distribution of a uniform random variable on $[0, 1]$ and the proof of the 2.3.2 follows immediately using the same argument used to prove inequality (2.14) in Theorem 2.2.2. □

2.3.3 Application to the coarse-to-fine algorithm

Subsection 2.3.2 provided us with an estimator $\hat{J} = \hat{J}_\varepsilon$ in (2.20) such that $\hat{J} > J$ with probability larger than $1 - \varepsilon$, which implies that

$$\text{FWER}(\hat{A}) \leq |V|\, p_{00}(\theta_G, \theta_V) + \hat{J}\, p_0(\theta'_V),$$

with probability $1 - \varepsilon$ at least.

In section 2.2, we provided a nonparametric coarse-to-fine procedure controlling the FWER by choosing constants θ_G, θ_V and θ'_V controlling the upper-bound at a

significance level α. This was done using a deterministic upper-bound of J, but cannot be directly applied with a data-based estimation of J because this would define data-dependent constants, which cannot be plugged into the definition of the set $\hat{A}$ without invalidating our estimation of the FWER. In other terms, if, for a fixed number J', one defines $\hat{A}_{J'}$ to be the discovery set obtained by optimizing θ_G and θ_V subject to $|V|\, p_{00}(\theta_G, \theta_V) + J'\, p_0(\theta'_V) \leq \alpha$, our previous results imply that $\mathrm{FWER}(\hat{A}_{J'}) \leq \alpha$ for all $J' \geq J$, but not necessarily that $\mathrm{FWER}(\hat{A}_{\hat{J}}) \leq \alpha + \varepsilon$.

A simple way to address this issue is to replace $\hat{A}_{\hat{J}}$ with

$$\tilde{A} = \bigcap_{J' \leq \hat{J}} \hat{A}_{J'}.$$

Because $\tilde{A} \subset \hat{A}_J$ with probability at least $1 - \varepsilon$, we have

$$\mathrm{FWER}(\tilde{A}) = P(\tilde{A} \cap V_0 \neq \emptyset) \leq P(\hat{A}_J \cap V_0 \neq \emptyset) + \varepsilon = \mathrm{FWER}(\hat{A}_J) + \varepsilon,$$

so that $\tilde{A}$ controls the FWER at level $\alpha + \varepsilon$ as intended.

2.3.4 Suggested Coarse-to-fine Procedure

The coarse-to-fine estimator relies on the choices of the constants θ_G, θ_V, θ'_V and ε_G and on the number of simulations, K. They were determined as follows in our experiments, for a control of the FWER at level α.

i. Fix $\epsilon < \alpha$ and compute $\hat{J}_\epsilon$, the estimated upper bound of J. We took $\epsilon = \frac{\alpha}{10}$ in our experiments.

ii. Fix a small $\delta > 0$ ($\delta = 10^{-4}$ in our experiments), and select $\theta_V = \theta_G$ and $\theta'_V = |V|\theta_G^2/\hat{J}_\epsilon$ such that $2|V|\theta_G^2 \leq \alpha - \epsilon - \delta$.

iii. We choose any K and ε_G such that $|V|c_K(\theta_G, \varepsilon_G) \leq \delta$, for some small $\delta > 0$.

These choices, which have the merit to be simple, albeit non-optimal, were found to perform well in our simulations (see section 2.5).

2.4 Model-based Analysis

In this section, we propose an alternate coarse-to-fine testing procedure, adapted to a specific regression model. In this framework, it is possible to obtain estimates for the power of the obtained test, and optimize its parameters on this basis. We will use this analysis as a benchmark to compare with the general non-parametric approach provided in the previous sections. We will assume, for simplicity, that all the cells have the same size, so $|g|$ is constant for $g \in G$.

2.4.1 Regression model

We assume that the observation is a realization of an i.i.d. family of random variables $U = ((Y^k, X^k), k = 1, \ldots n)$ where the Y's are real-valued and $X^k = (X_v^k, v \in V)$

is a high-dimensional family of variables indexed by the set V. We also assume that X_v^k, $v \in V$, are independent and centered Gaussian, with variance σ_v^2, and that

$$Y^k = a_0 + \sum_{v \in A} a_v X_v^k + \psi^k$$

where $\psi^1, \ldots, \psi^n$ are i.i.d. Gaussian with variance σ^2, and a_v, $v \in A$, are unknown real coefficients. We will denote by $\mathbf{Y}$ the vector $(Y^1, \ldots, Y^n)$ and let $\bar{\mathbf{Y}} = \left(\sum_{k=1}^n Y^k/n\right) \mathbf{1}_n$ where $\mathbf{1}_n$ is the vector composed by ones repeated n times. We also let $\mathbf{X}_v = (X_v^1, \ldots, X_v^n)$ and $\boldsymbol{\psi} = (\psi^1, \ldots, \psi^n)$, so that

$$\mathbf{Y} = \sum_{v \in A} a_v \mathbf{X}_v + \boldsymbol{\psi}.$$

Finally, we will denote by $\sigma_{\mathbf{Y}}^2$ the common variance of $Y^1, \ldots, Y^n$ and assume that it is known (or estimated from the observed data).

2.4.2 Scores

For $v \in V$, we denote by π_v the orthogonal projection on the subspace S_v spanned by the two vectors $\mathbf{X}_v$ and $\mathbf{1}_n$. We will also denote by π_g ($g \in G$) the orthogonal

projection on the subspace S_g spanned by the vectors $\mathbf{X}_v$, $v \in g$, and $\mathbf{1}_n$, and let

$$T_g(U) = \frac{\|\pi_g \mathbf{Y}\|^2 - \|\bar{\mathbf{Y}}\|^2}{\sigma^2_{\mathbf{Y}}},$$
$$T_v(U) = \frac{\|\pi_v \mathbf{Y}\|^2 - \|\bar{\mathbf{Y}}\|^2}{\sigma^2_{\mathbf{Y}}}.$$

(The projections are simply obtained by least-square regression of $\mathbf{Y}$ on $\mathbf{X}_v, v \in g$, for π_g and on $\mathbf{X}_v$ for π_v.) We now provide estimates of

$$p_{00}(\theta_G, \theta_V) = \mathbb{P}\left(\frac{\|\pi_g \mathbf{Y}\|^2 - \|\bar{\mathbf{Y}}\|^2}{\sigma^2_{\mathbf{Y}}} > \theta_G; \frac{\|\pi_v \mathbf{Y}\|^2 - \|\bar{\mathbf{Y}}\|^2}{\sigma^2_{\mathbf{Y}}} > \theta_V\right)$$

for $v \in V_{00}$ and $g = g(v)$ and

$$p_0(\theta_V) = \mathbb{P}\left(\frac{\|\pi_v \mathbf{Y}\|^2 - \|\bar{\mathbf{Y}}\|^2}{\sigma^2_{\mathbf{Y}}} > \theta_V\right)$$

for $v \in V_0$. Note that, because we consider residual sums of squares, we here use large values of the scores in the rejection regions (instead of small values in the introduction and other parts of the paper), hopefully without risk of confusion.

Proposition 2.4.1. *For all θ_G and θ_V and $g \in G_0$:*

$$p_{00}(\theta_G, \theta_V) \leq C(|g|) \exp\left(-\frac{\theta_G}{2}\right) \theta_G^{\frac{|g|}{2}} \left(1 - G_\beta\left(\frac{\theta_V}{\theta_G}, \frac{1}{2}, \frac{|g|+1}{2}\right)\right) + (1 - F_1(\theta_G - |g| + 1)),$$

where F_1 is the c.d.f. of a chi-squared distribution with one degree of freedom and

$$C(|g|) = \frac{\exp\left(\frac{|g|-1}{2}\right)}{\sqrt{2}(|g|-1)^{\left(\frac{|g|-1}{2}\right)}} \frac{\Gamma(\frac{|g|}{2}+\frac{1}{2})}{\Gamma(\frac{|g|}{2}+1)}.$$

Moreover

$$p_0(\theta_V) \leq 1 - F_1(\theta_V).$$

Note that the upper-bound for p_{00} is larger than 1 when $\theta_G \leq |g| - 1$, so that this estimate is useful only when $\theta_G > |g| - 1$.

Proof. For $v \in V_{00}$ and $g = g(v)$, we have

$$\sigma^2_{\mathbf{Y}} = \sum_{v \in A} a_v^2 \sigma_v^2 + \sigma^2 = \sum_{v \in A \cap g^c} a_v^2 \sigma_v^2 + \sigma^2$$

because $A \cap g^c = A$. Consider the conditional probability:

$$\mathbb{P}\left(\frac{\|\pi_g \mathbf{Y}\|^2 - \|\bar{\mathbf{Y}}\|^2}{\sigma^2_{\mathbf{Y}-g}} > \theta_G; \frac{\|\pi_v \mathbf{Y}\|^2 - \|\bar{\mathbf{Y}}\|^2}{\sigma^2_{\mathbf{Y}-g}} > \theta_V \,\middle|\, (\mathbf{X}_v)_{v \in g} \right).$$

The conditional distribution of $\mathbf{Y}$ given $(\mathbf{X}_v)_{v \in g}$ is Gaussian $\mathcal{N}(0, \sigma^2_{\mathbf{Y}}\, I_n)$ (where I_n is the n-dimensional identity matrix). Denote by π'_v the projection on the orthogonal complement of $\mathbf{J}$ in S_v and by π'_g the projection on the orthogonal complement of S_v in S_g, so that

$$\|\pi_g \mathbf{Y}\|^2 - \|\bar{\mathbf{Y}}\|^2 = \|\pi'_g \mathbf{Y}\|^2 + \|\pi'_v \mathbf{Y}\|^2$$

and

$$\|\pi_v \mathbf{Y}\|^2 - \|\bar{\mathbf{Y}}\|^2 = \|\pi'_v \mathbf{Y}\|^2.$$

This implies that:

$$\mathbb{P}\left(\frac{\|\pi_g \mathbf{Y}\|^2 - \|\bar{\mathbf{Y}}\|^2}{\sigma^2_{\mathbf{Y}}} > \theta_G; \frac{\|\pi_v \mathbf{Y}\|^2 - \|\bar{\mathbf{Y}}\|^2}{\sigma^2_{\mathbf{Y}}} > \theta_V \,\middle|\, (\mathbf{X}_v)_{v\in g} \right) = \mathbb{P}\left(\frac{\|\pi'_g \mathbf{Y}\|^2 + \|\pi'_v \mathbf{Y}\|^2}{\sigma^2_{\mathbf{Y}}} > \theta_G; \frac{\|\pi'_v \mathbf{Y}\|^2}{\sigma^2_{\mathbf{Y}}} > \theta_V \,\middle|\, (\mathbf{X}_v)_{v\in g} \right)$$

At this stage, one can apply Cochran's theorem to $P'_g(\mathbf{Y}/\sigma_{\mathbf{Y}})$ and $P'_v(\mathbf{Y}/\sigma_{\mathbf{Y}})$, which are conditionally independent given $\mathbf{X}_v$, $v \notin G$, to reduce the problem to finding an upper bound for:

$$\mathbb{P}\left(\eta + \zeta \geq \theta_G; \zeta \geq \theta_V\right),$$

where η is $\chi^2(|g| - 1)$ and ζ is $\chi^2(1)$, and the two variables are independent.

Assume that $\theta'_G = \theta_G - |g| + 1 > 0$ and write this probability as

$$\mathbb{P}\left(\eta + \zeta \geq \theta_G, \theta_V \leq \zeta < \theta'_G\right) + \mathbb{P}\left(\eta + \zeta \geq \theta_G, \zeta \geq \theta_V, \zeta \geq \theta'_G\right),$$

which is less than:

$$\mathbb{P}\left(\eta + \zeta \geq \theta_G, \theta_V \leq \zeta < \theta_G\right) + (1 - F_1(\theta'_G)).$$

The first term in the sum can be re-written as

$$\mathbb{E}\left(\mathbb{P}(\eta \geq \theta_G - \zeta|\zeta)\mathbf{1}_{\theta_V \leq \zeta < \theta'_G}\right).$$

We will use the following tail inequality for the c.d.f., F_k, of a $\chi^2(k)$ random variable, stating that

$$1 - F_k(zk) \leq (z\exp(1-z))^{\frac{k}{2}},$$

for any $z > 1$. We apply this result to $k = |g| - 1$ and $z = \frac{\theta_G - V}{|g|-1}$ and write (using the fact that the p.d.f. of a $\chi^2(1)$ is $z^{-1/2}e^{z/2}/(\sqrt{2}\Gamma(1/2))$ for $z > 0$)

$$\begin{aligned}
\mathbb{E}\left(\mathbb{P}(\eta \geq \theta_G - \zeta|\zeta)\mathbf{1}_{\theta_V \leq \zeta < \theta'_G}\right) &\leq \mathbb{E}\left(\left(\frac{\theta_G - \zeta}{|g|-1}\exp\left(1 - \frac{\theta_G - \zeta}{|g|-1}\right)\right)^{\frac{|g|-1}{2}}\mathbf{1}_{\theta_V \leq \zeta < \theta_G}\right) \\
&= \int_{\theta_V}^{\theta_G} \frac{(\theta_G - z)^{|g|/2-1/2}}{\Gamma(1/2)\sqrt{2}(|g|-1)^{|g|/2-1/2}} z^{-1/2}e^{-\frac{tg'}{2}}dz \\
&\leq \frac{\theta_G^{\frac{|g|}{2}}e^{-\frac{\theta'_G}{2}}}{\Gamma(1/2)\sqrt{2}(|g|-1)^{|g|/2-1/2}}\int_{\theta_V/tg}^{1}(1-z)^{|g|-1/2}z^{-1/2}dz \\
&= C(|g|)\theta_G^{\frac{|g|}{2}}e^{-\frac{\theta_G}{2}}\left(1 - G_\beta\left(\frac{\theta_V}{\theta_G}, \frac{1}{2}, \frac{|g|+1}{2}\right)\right).
\end{aligned}$$

The second upper-bound, for $p_0(\theta_V)$, is easily obtained, and left to the reader. □

This leads us immediately to the following corollary:

Corollary 2.4.1. *With the thresholds θ_G and θ_V, an upper bound of the FWER is:*

$$\mathrm{FWER}(\hat{A}) \leq |V| C(|g|) \exp\left(-\frac{\theta_G}{2}\right) \theta_G^{\frac{|g|}{2}} \left(1 - G_\beta\left(\frac{\theta_V}{\theta_G}, \frac{1}{2}, \frac{|g|+1}{2}\right)\right) + |V| \left(1 - F_1(\theta_G - |g| + 1)\right) + J|g| \left(1 - F_1(\theta_V)\right). \quad (2.23)$$

Figure 2.2 provides an illustration of the level curves associated to the above FWER upper bound. More precisely, it illustrates the tradeoff between the conservativeness at the cell level and at the individual index level. In the next section, the optimization for power will be made along these level lines. Figure 1 also provides the value of the Bonferroni-Holm threshold. For the coarse-to-fine procedure to be less conservative than the Bonferroni-Holm approach, we need the index-level threshold to be smaller, i.e., the optimal point on the level line to be chosen below the corresponding dashed line.

The derivation of (2.23) is based on the assumption that we have a fixed cell size (across all the cells). If this is not true, it is easy to generalize the previous upper bound. Letting

$$\phi(|g|, \theta_G, \theta_V) = C(|g|) \exp\left(-\frac{\theta_G}{2}\right) \theta_G^{\frac{|g|}{2}} \left(1 - G_\beta\left(\frac{\theta_V}{\theta_G}, \frac{1}{2}, \frac{|g|+1}{2}\right)\right),$$

it suffices to replace $|V|\phi(|g|, \theta_G, \theta_V)$ in (2.23) with $\sum_{g \in G} |g| \phi(|g|, \sqrt{|g|}\theta_G, \theta_V)$ where θ_G does not depend on the cell g.

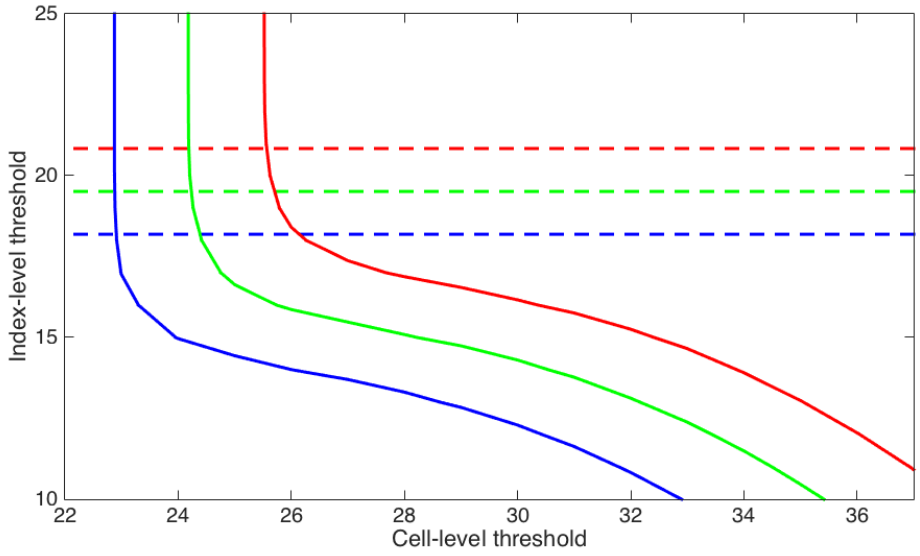

Figure 2.2: Level curves of the upper bound of the FWER for the levels 0.2 (blue), 0.1 (green) and 0.05 (red). The horizontal dashed lines represent the thresholds at the individual level for a Bonferroni-Holm test, with corresponding colors. For this figure, $V = 10^4$, $J = 600$ and $g = 10$

2.4.3 Optimal thresholds

Equation (2.23) provides a constraint on the pair (θ_G, θ_V) to control the FWER at a given level. We now show how to obtain "optimal" thresholds (θ_G^*, θ_V^*) that maximize the probability of detection subject to this constraint. The discussion will also help understanding how active indices clustering in cells improves the power of the coarse-to-fine procedure.

The conditional distribution of $\mathbf{Y}$ given $(\mathbf{X}_v, v \in g)$ is $\mathcal{N}(\sum_{v \in g \cap A} a_v \mathbf{X}_v, \sigma^2_{\mathbf{Y}-g})$ with $\sigma^2_{\mathbf{Y}-g} = \sum_{v \in A \cap g^c} a_v^2 \sigma_v^2 + \sigma^2$. It follows from this that, conditionally to these variables, $(\|P_g \mathbf{Y}\|^2 - \|\bar{\mathbf{Y}}\|^2)/\sigma^2_{\mathbf{Y}-g}$ follows a non-central chi-square distribution $\chi^2(\rho_g(\mathbf{X}_v, v \in$

$g), |g|)$, with

$$\rho_g(\mathbf{X}_v, v \in g) = \frac{\|\sum_{v \in g \cap A} a_v(\mathbf{X}_v - \bar{X}_v)\|^2}{\sigma^2_{\mathbf{Y}-g}}$$

where $\bar{X}_v = \frac{1}{n}\sum_{k=1}^{n} X_v^k \mathbf{1}_n$. Using the fact that $\rho_g(\mathbf{X}_v, v \in g)/n$ converges to

$$\rho_g := \frac{\sum_{v \in g \cap A} a_v^2 \sigma_v^2}{\sigma^2_{\mathbf{Y}-g}},$$

we will work with the approximation

$$\frac{\|P_g \mathbf{Y}\|^2 - \|\bar{\mathbf{Y}}\|^2}{\sigma^2_{\mathbf{Y}-g}} \sim \chi^2(n\rho_g, |g|).$$

With a similar analysis, and letting for $v \in A$, $\sigma^2_{\mathbf{Y}-v} = \sum_{v' \in A \setminus v} a^2_{v'}\sigma^2_{v'} + \sigma^2$, we will assume that

$$\frac{\|P_v \mathbf{Y}\|^2 - \|\bar{\mathbf{Y}}\|^2}{\sigma^2_{\mathbf{Y}-v}} \sim \chi^2(n\rho_v, 1)$$

with

$$\rho_v := \frac{a_v^2 \sigma_v^2}{\sigma^2_{\mathbf{Y}-v}}.$$

Therefore, an approximation of a lower bound for the probability of detection of an active index v in a cell g will be:

$$\mathbb{P}\left(v \in \hat{A}\right) \geq 1 - F_{|g|}(\theta_G, n\rho_g) - F_1(\theta_V, n\rho_v), \tag{2.24}$$

where $F_k(x, \delta)$ is the c.d.f of a non-central chi-squared distribution with k degrees of

freedom and δ as a non-centrality parameter evaluated at x.

We use the lastest result in the following way. One can fix a target effect size η (the ratio of the effect of $\mathbf{X}_v$ compared to the total variance of $\mathbf{Y}$), and a target cluster size, k, that represents the number of active loci that we expect to find in an active cell, and take $\rho_v = \eta$ and $\rho_g = k\eta$ to optimize the lower-bound in (2.24) subject to the FWER constraint (2.23). This provides optimal constants (θ_G, θ_V) for this target case. This is illustrated with numerical simulations in the next section.

2.5 Simulations and power comparison

In this section, we will first generate simulations under the model of section 2.4. The purpose of the simulations will be to first show the effect of the coarse-to-fine algorithm on the detection power. It will also illustrate the effect of optimizing the thresholds, assuming a parametric model. Of course, as mentioned in the introduction, the default coarse-to-fine algorithm that should be considered is the non-parametric version of section 2.2. Therefore, we will compare it with the Bonferroni-Holm approach but also the parametric coarse-to-fine when the data has been generated by the parametric model.

Our second set of experiments uses the software PLINK[36] to simulate case control genome-wide association studies, where the indices will corresponds to SNPs and compare the (non-parametric) coarse-to-fine approach to the Bonferroni-Holm procedure.

In this setting we will also allow the variables X_v to be correlated.

As a main expected observation, all these simulations will illustrate the idea that the more clustered the active indices, the more powerful the coarse-to-fine procedure will be compared to the Bonferroni-Holm procedure.

2.5.1 Simulations under the parametric model

We let $|V| = 10^4$ with 400 cells of size $|g| = 25$ each. We assume $n = 300$ observations for the model

$$Y^k = a_0 + \sum_{v \in A} a_v X_v^k + \psi^k, k = 1, \ldots, n$$

with $a_0 = 0$ and $a_v = 1$, for all $v \in A$. We also let the X_v's and ψ be i.i.d standard normals. We will control the FWER at level $\alpha = 0.1$.

We will consider two versions of the parametric coarse-to-fine procedure. The first one is a best-case scenario, run under the optimistic assumption that the true values of ρ_g and ρ_v are known in (2.24). The second is a more realistic, but sub-optimal, procedure in which the sum of the first two terms in (2.23), and the last term in the same equation are adjusted to both equal $\alpha/2$. Both will be compared to the Bonferroni-Holm procedure.

The first simulation illustrates the effect of optimization over the thresholds on the parametric version of the coarse-to fine algorithm. More precisely, we consider

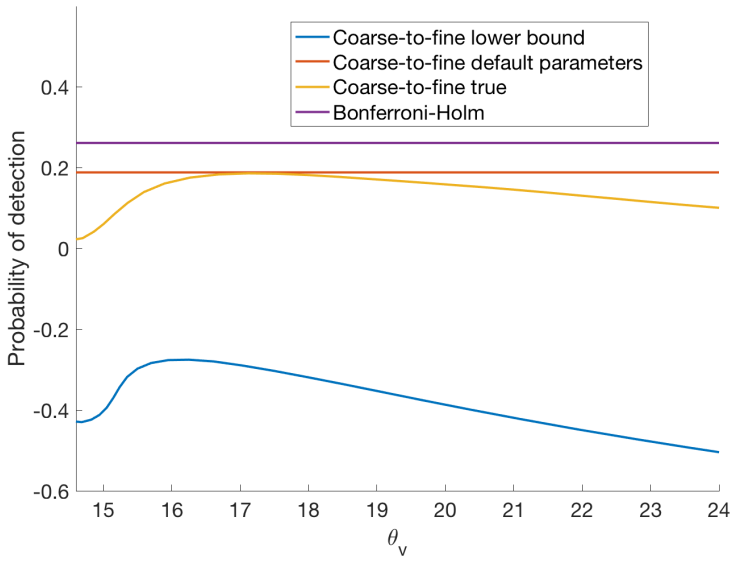

Figure 2.3: Probability of detection as a function of θ_V in the admissible space, using the parametric coarse-to-fine procedure for an active cell with 1 active index. The value of θ_G is determined by the implicit equation $FWER(\theta_V, \theta_G) = \alpha$. Coarse-to-fine true represents the estimated true probability of detection via Monte Carlo simulation. Coarse-to-fine lower bound represents the lower bound of the probability of detection obtained via (2.24). We fixed $\hat{J}_{0.01}$ to the value 40×25, which is an upper bound of $\hat{J}$ for all the simulations performed. As expected, the Bonferroni-Holm procedure is better, given that the clustering assumption is not true.

3 scenarios of a fixed active index contained in a cell having respectively 0,1 and 2 other active indices. For each of these scenarios, we will be interested in the probability of detection of such an index. This is illustrated in figures 2.3, 2.4 and 2.5. Unsurprisingly, the procedure using optimized parameters outperforms the other two but the coarse-to-fine approach using default parameters significantly improves on Bonferroni-Holm when the number of active indices in the cell is more than 1. We also note that the lower bound computed in (2.24) is most of the time quite close to the true probability of detection.

In the second set of simulations, we consider five scenarios varying the number of

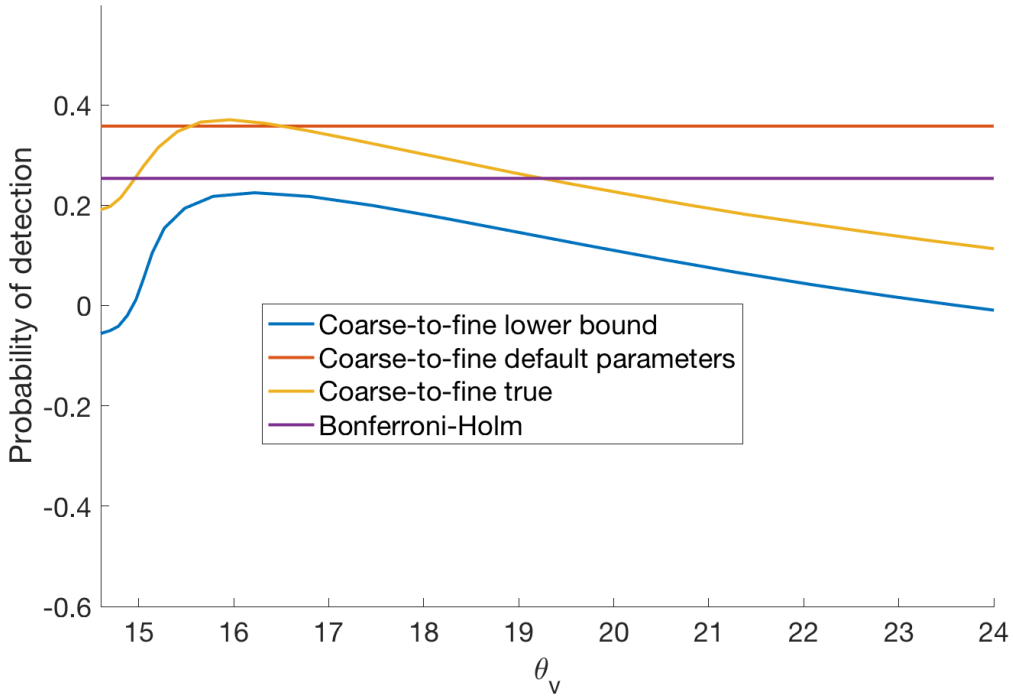

Figure 2.4: Probability of detection as a function of θ_V in the admissible space, using the parametric coarse-to-fine procedure for an active cell with 2 active indices. The CTF procedure outperforms Bonferroni-Holm in this case (even when using the default choice for the thresholds). See Fig. 2.3 for additional details.

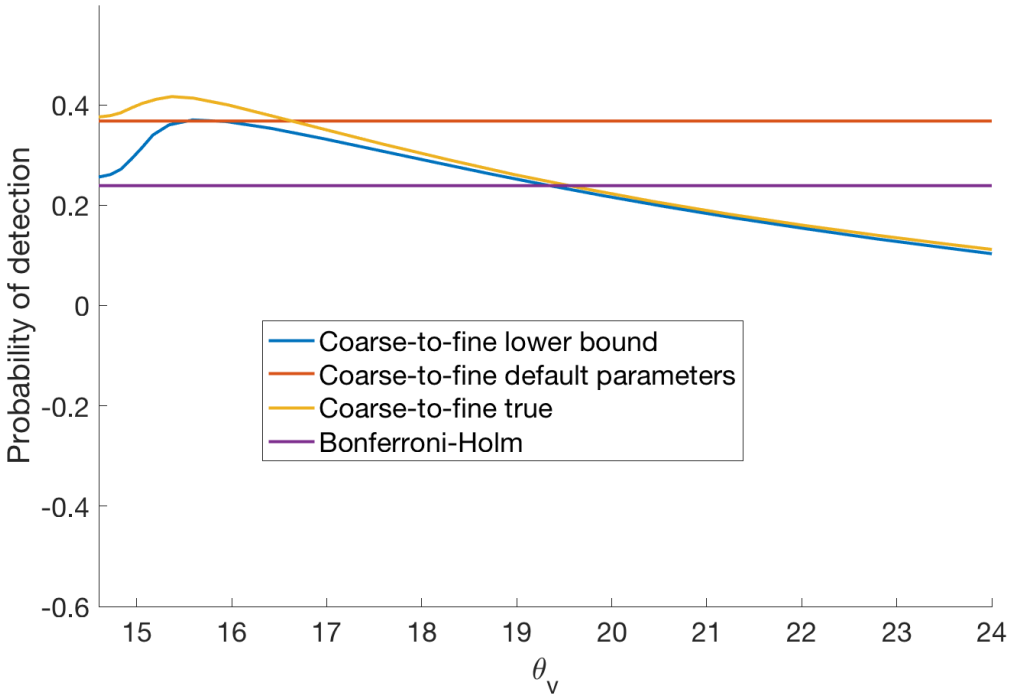

Figure 2.5: Probability of detection as a function of θ_V in the admissible space, using the parametric coarse-to-fine procedure for an active cell with 3 active indices. The CTF procedure outperforms Bonferroni-Holm. See Fig. 2.3 for additional details.

Active indices per active cell	Optimal CTF	Parametric CTF	Nonparametric CTF	BH
1	3.88	3.88	3.85	5.11
2	7.82	7.35	6.09	5.11
4	9.69	8.56	8.13	5.11
5	11.03	8.68	8.15	5.11
20	11.52	9.34	9.12	5.11

Table 2.1: Average number of true detections for each of the 4 methods, from left to right: coarse-to-fine using the optimized thresholds, parametric coarse-to-fine using the default thresholds, non parametric coarse-to-fine using default parameters, and Bonferoni-Holm. The total number of active indices is 20 in all cases.

active cells and indices, namely (1) 20 active cells with 1 active index each; (2) 10 active cells with 2 active indices; (3) 4 active cells with 5 active indices; (4) 2 active cells with 10 active indices; (5) 1 active cell with 20 active indices. In each case, we ran 100 simulations from which we computed the average number of true detections. The results are provided in table 2.1 and also include the non-parametric coarse-to-fine method. We found that the fully-informed parametric methods outperforms all others with some margin, the parametric method with default parameters is only slightly better than the non-parametric one. All three outperform Bonferroni-Holm as soon as the number of active indexes in cells is more than 1.

Figure 2.6 provides the average estimated upper-bound for the number of active cells used in the coarse-to-fine methods. Even is this upper-bound is conservative and estimate about 20 more cells that their real number, the number of detections is only slightly affected.

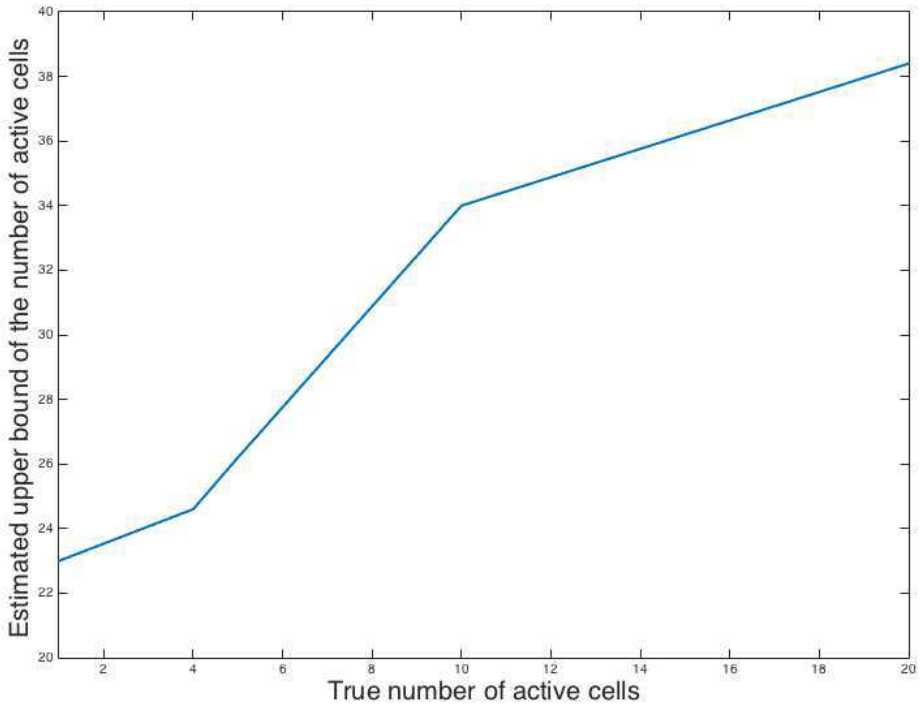

Figure 2.6: Plot of the average upper bound of the number of active cells as a function the true number. Even though this upper bound is not particularly tight, it will be sufficient to ensure that coarse-to-fine outperforms the Bonferroni-Holm procedure.

2.5.2 Simulations using the PLINK software

In a second set of simulations, we use the PLINK software to generate case control studies, where the indexes $v \in V$ represent SNPs. The variable X_v takes ternary values: 0 if both alleles in the SNP are wild-type (the major allele in the population), 1 if one of the alleles is a variant and 2 if both alleles are. The major allele frequency range is $[0.8, 0.95]$. The total number of SNPs is $|V| = 10^4$. The X_v's will either be simulated as independent variables, of with some "linkage disequilibrium" (LD) in which case each SNP is paired is another with a correlation equal to 0.8.

From these SNPs, a binary phenotype Y (cases vs. controls) is generated, yielding $n = 600$ samples, 300 cases ($Y = 1$) and 300 controls ($Y = 0$). The generative model for Y is logistic

$$P(Y = 1|X_v, v \in V) \propto \exp\left(a_0 + \sum_{v \in A} a_v X_v Y\right),$$

with $a_v = \log 2$ for $v \in A$. This sets the odds ratio for active SNPs is set to 2, where

$$\text{odds ratio} = \frac{P(Y = 1|X_v = 1, X_{v'}, v' \neq v)/P(Y = 0|X_v = 1, X_{v'}, v' \neq v)}{P(Y = 1|X_v = 0, X_{v'}, v' \neq v)/P(Y = 0|X_v = 0, X_{v'}, v' \neq v)} = e^{a_v}.$$

We consider cells (loosely interpreted as "genes") of fixed size, ν_G, with a random assignment of active SNPs to cells based on a variant of a Chinese restaurant process.[37] More precisely, assume that the active indices are $1, 2, ..., |A|$, and denote

by $C_1, C_2, \ldots, C_{|A|}$ the random variables representing the cells to which each active index is assigned. The sequence $C_1, \ldots, C_{|A|}$ is generated as follows.

i) C_1=1

ii) Iterate over $k = 1, 2, \ldots, |A|$ ($|A| = 25$ for all cases). For a given k, let $N_k = \max_k\{C_1, \ldots, C_k\}$ and for $i = 1, \ldots, N_k$, let $n_i = \sum_{j=1}^{k} \mathbf{1}_{C_j=i}$ be the number of indices assigned to cell i. Then

$$\boldsymbol{P}\left(C_{k+1} = i\right) \propto \begin{cases} \dfrac{\alpha}{k+\alpha} \text{ if } i = N_k + 1 \\ \dfrac{n_i}{k+\alpha} \text{ if } i \leq N_k \text{ and } n_i < \nu_G \\ \quad 0 \text{ otherwise} \end{cases}$$

Here, α is a parameter controlling the clustering of the indices. The smaller α, the more clustered the active indices will be within active cells (see figure 2.7).

We generated datasets with the previous parameters, iterating over $\alpha = 0.5$, 1, 5, 10, 20, 30, 40 and 50 in the Chinese restaurant process and considering four cases: (I) $\nu_G = 10$, no LD; (II) $\nu_G = 25$, no LD; (III) $\nu_G = 10$, LD = 0.8; (III) $\nu_G = 25$, LD = 0.8. In each case, we took the average over 50 simulations.

Since we are interested in the effect of clustering on the performance of the coarse-to-fine algorithm compared to the Bonferroni-Holm procedure, and not the effect of α itself, we excluded the rare events where we generated a random clustering where the

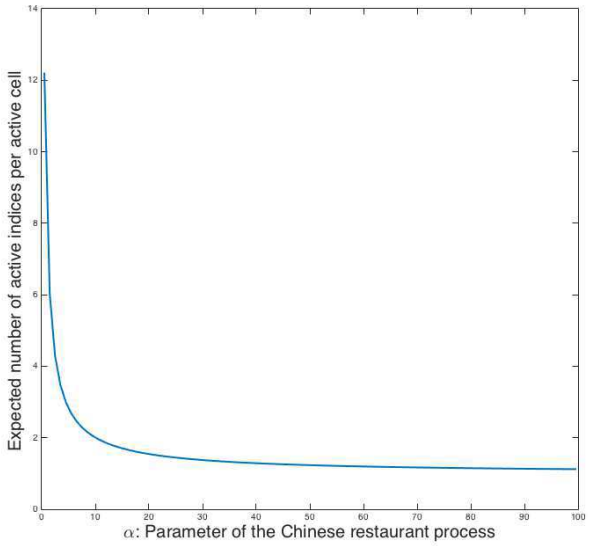

Figure 2.7: Expected number of active indices per active cell as a function of α, the clustering parameter of our assignment process. In this case where the size of a cell is greater or equal than the number of active indices, our clustering process corresponds exactly to a Chinese restaurant process.

number of active cells decreased after increasing the parameter α. Table 2.2 illustrates the estimation of the upper bound of the number of active cells for the independent and correlated datasets. Tables 2.3 compares the performance of the non-parametric coarse-to-fine procedure with the Bonferroni-Holm procedure.

	Independent SNPs				Correlated SNPs			
	$\|g\| = 25$		$\|g\| = 10$		$\|g\| = 25$		$\|g\| = 10$	
α	J	$\hat{J}$	J	$\hat{J}$	J	$\hat{J}$	J	$\hat{J}$
0.5	50	595	30.4	344.2	85	535	34.2	184.6
1	132.5	658	51.2	373.8	128	556.5	50.6	192
5	232	711	94	398.4	218.5	600	89.8	217.2
10	316	754.5	124.4	426.6	298	632.5	119.4	232.8
20	390.5	791	154.4	447.2	376	690	151.2	255.6
30	454	809	178.8	463.2	441	704.5	177.6	271.2
40	502	821.5	192.6	481	500.5	730.5	197	282
50	530.5	830	202.2	482.4	525	750	212.2	285.8

Table 2.2: Comparison between the true number of indices in active cells (J) and the estimated upper bound ($\hat{J}$) averaged over 50 simulations, as a function of the clustering parameter of the Chinese restaurant process (α), for cell sizes $|g| = 25$ or 10, in the independent and correlated cases.

	Independent SNPs				Correlated SNPs			
	$\|g\| = 25$		$\|g\| = 10$		$\|g\| = 25$		$\|g\| = 10$	
α	CTF	BH	CTF	BH	CTF	BH	CTF	BH
0.5	10.6	5.2	9.4	5.08	10.6	4.76	11.6	3.7
1	9.94	5.2	8.66	5.08	10.14	4.76	10.08	3.7
5	9.18	5.2	7.76	5.08	9.52	4.76	10.02	3.7
10	8.14	5.2	7.36	5.08	8	4.76	9.68	3.7
20	7.22	5.2	7.04	5.08	7.28	4.76	8.24	3.7
30	6.26	5.2	6.88	5.08	6.68	4.76	7.64	3.7
40	4.06	5.2	5.2	5.08	5.06	4.76	5.8	3.7
50	3.8	5.2	5	5.08	4.6	4.76	4.98	3.7

Table 2.3: Comparison between the average number of true detections for the coarse-to-fine (CTF) and Bonferroni-Holm (BH) procedures averaged over 50 suimulations, as functions of the clustering parameter of the Chinese restaurant process (α), for cell sizes $|g| = 25$ or 10, in the independent and correlated cases.

Chapter 3

Dependence adapted coarse-to-fine procedure controlling the FWER

Even if the coarse-to-fine procedure derived in the previous chapter controls the FWER for any dependence structure between the tests, it is well suited for the situation where the test statistics $T_v(\mathbf{U})$ are independent. To see this, imagine the extreme situation where all the statistics $T_v(\mathbf{U})$ are identical: $T_v(\mathbf{U}) = T_{v'}(\mathbf{U})$ for any v and v', and that all null hypotheses are false. The Bonferroni bound will reject all the hypotheses if and only if one of the T_v's is less than $\frac{\alpha}{|V|}$. Otherwise, the detection set will be empty. As for our coarse-to-fine strategy, using proposition 2.1.1, one can see that a necessary condition to have a non empty detection set is that $T_v(\mathbf{U}) \leq \frac{\alpha}{\hat{J}}$. Moreover, our method for computing $\hat{J}$ is likely to have the following issue in this situation: all the statistics will be less than the conservative threshold t_0, and $\hat{J}$ will

simply be the total number of tests $|V|$. As a consequence, the procedure will be at least as conservative as the Bonferroni procedure. Imagine now an "oracle" testing procedure having the information that all the statistics are identical. The procedure therefore knows that it is sufficient to set α as a threshold instead of $\frac{\alpha}{|V|}$, which is a considerable improvement. In the general case, if the rejection region is the same for every hypothesis $v \in V$ and is denoted by Γ, the FWER can be written as following:

$$\mathbb{P}\left(\min_{v\in V_0}(T_v(\mathbf{U}), v \in V_0) \in \Gamma\right).$$

It is then sufficient for the oracle procedure to know the distribution of the statistic $\min_{v\in V_0}(T_v(\mathbf{U}), v \in V_0)$. The idea behind the procedure described at the end of section 1.6 addresses this very issue via estimating in a conservative way that distribution. The purpose of this chapter is to derive a coarse-to-fine procedure in the same spirit. The high level idea consists in controlling an upper bound of the FWER that is sharper than the Bonferroni bound in a setting where the tests are highly dependent. This is done by just writing the FWER as a union of events and avoid breaking this union and replacing them by a sum of probabilities. Similarly to the previous chapter, we will first present a procedure assuming asymptotic resampling. In a second part we will propose 3 different ways of modifying the asymptotic resampling procedure in order to have a finite resampling procedure controlling the FWER. Finally, we will first propose simulations under a toy example in order to see the effect of dependence

between tests on different procedures. We will then present simulations comparing all the procedures presented so far using PLINK under a setting identical to the previous chapter's simulations.

3.1 Dependence adapted coarse-to-fine procedure assuming asymptotic resampling

As announced, we start by presenting the procedure assuming asymptotic resampling. Before presenting the algorithm, let us first set notations, and the assumptions under which the FWER will be controlled.

3.1.1 Notations and assumptions

Let I be an interval of $\mathbb{R}$. In this setting, we will consider a family of triples parametrized by a real number $t \in I$, denoted by $(\theta_G(t), \theta_V(t), \epsilon(t))$. The parametrization will satisfy the following conditions:

A1. The function $t \to \theta_V(t) - \epsilon(t)$ is nondecreasing.

A2. The function $t \to \theta_G(t)$ is nondecreasing.

A3. The function $t \to \epsilon(t)$ is nondecreasing

A4. The function $t \to (\theta_G(t), \theta_V(t), \epsilon(t))$ is left continuous.

In the following procedure, $\theta_G(t)$ and $\theta_V(t)-\epsilon(t)$ will correspond to the thresholds at the cell level and at the index level, and the scores $T_v(\mathbf{U})$ and $T_g(\mathbf{U})$ defined respectively for every $v \in V$ and $g \in G$ can be any scores. The assumptions [A1.] and [A2.] will therefore ensure that the detection set will get larger as t increases, since we will reject the index when the scores at the corresponding cell level and the index level will be less than these thresholds. Assumptions [A3.] and [A4.] (together with the two first assumptions) are made in order to satisfy the following lemma that will be used to show that our procedure will control the FWER at a given level α.

Lemma 3.1.1. *Denote by $R(\mathbf{U}, \theta_G)$ the set: $\{v \in V : T_{g(v)}(\mathbf{U}) < \theta_G\}$. Let V_1, V_2 and V_3 be any subsets of V. Denote by:*

$$\Delta(\mathbf{U}, V_2, V_3, \theta_G) = \min_{v \in V_2 \cap R(\mathbf{U}, \theta_G)} T_v(\mathbf{U}) - \min_{v \in V_3} T_v(\mathbf{U}).$$

Then:

$$t \to \mu\left(\xi : \bigcup_{v \in V_1} \big(T_v(\xi \odot \mathbf{U}) < \theta_V(t), T_{g(v)}(\xi \odot \mathbf{U}) < \theta_G(t)\big) \bigcup \Delta(\xi \odot \mathbf{U}, V_2, V_3, \theta_G(t)) \geq \epsilon(t)\right)$$

is left continuous.

Proof. We start by rewriting

$$\mu\left(\xi : \bigcup_{v \in V_1} \big(T_v(\xi \odot \mathbf{U}) < \theta_V(t), T_{g(v)}(\xi \odot \mathbf{U}) < \theta_G(t)\big) \bigcup \Delta(\xi \odot \mathbf{U}, V_2, V_3, \theta_G(t)) \geq \epsilon(t)\right)$$

as:

$$\frac{1}{|\mathfrak{S}|}\sum_{\xi\in\mathfrak{S}}\max\left(\max_{v\in V_1}(\mathbf{1}_{T_v(\xi\odot\mathbf{U})<\theta_V(t)}\mathbf{1}_{T_{g(v)}(\xi\odot\mathbf{U})<\theta_G(t)}),\mathbf{1}_{\Delta(\xi\odot\mathbf{U},V_2,V_3,\theta_G(t))\geq\epsilon(t)}\right).$$

It is therefore sufficient to prove the following:

1. for every $\xi\in\mathfrak{S}$ and every $v\in V_1$: $t\to\mathbf{1}_{T_v(\xi\odot\mathbf{U})<\theta_V(t)}$ and $t\to\mathbf{1}_{T_{g(v)}(\xi\odot\mathbf{U})<\theta_G(t)}$ are left continuous.

2. for every $\xi\in\mathfrak{S}$: $t\to\mathbf{1}_{\Delta(\xi\odot\mathbf{U},V_2,V_3,\theta_G(t))\geq\epsilon(t)}$ is left continuous.

The statement (1) is true because for every $v\in V_1$, $t\to\mathbf{1}_{T_v(\xi\odot\mathbf{U})<t}$ and $t\to\mathbf{1}_{T_{g(v)}(\xi\odot\mathbf{U})<t}$ are left continuous and $t\to\theta_V(t)$ and $t\to\theta_G(t)$ are left continuous nondecreasing. To prove the statement (2), we notice that $\mathbf{1}_{\Delta(\xi\odot\mathbf{U},V_2,V_3,\theta_G(t))\geq\epsilon(t)}$ is :

$$\mathbf{1}_{\min_{v\in V_2\cap R(\mathbf{U},\theta_G)}T_v(\xi\odot\mathbf{U})-\min_{v\in V_3}T_v(\xi\odot\mathbf{U})-\epsilon(t)\geq 0},$$

which in turn can be rewritten as:

$$\mathbf{1}_{\min_{v\in V_2}\left(T_v(\xi\odot\mathbf{U})+\mathbf{1}_{\{T_{g(v)}(\xi\odot\mathbf{U})\geq\theta_G(t)\}}-\min_{v\in V_3}T_v(\xi\odot\mathbf{U})-\epsilon(t)\right)\geq 0}.$$

Now, for each $v\in V_2$, the function $t\to T_v(\xi\odot\mathbf{U})+\mathbf{1}_{T_{g(v)}(\xi\odot\mathbf{U})\geq\theta_G(t)}-\min_{v\in V_3}T_v(\xi\odot\mathbf{U})-\epsilon(t)$ is non-increasing and left continuous, and so is:
$t\to\min_{v\in V_2}\left(T_v(\xi\odot\mathbf{U})+\mathbf{1}_{T_{g(v)}(\xi\odot\mathbf{U})\geq\theta_G(t)}-\min_{v\in V_3}T_v(\xi\odot\mathbf{U})-\epsilon(t)\right)$. We finally

notice that $t \to \mathbf{1}_{t\geq 0}$ is right continuous. Combining these two last fact, we conclude that $t \to \mathbf{1}_{\Delta(\xi\odot\mathbf{U},V_2,V_3,\theta_G(t))\geq\epsilon(t)}$ is left continuous, which ends the proof. □

3.1.2 Algorithm

We now describe the procedure and will show that under assumptions [A1,A2,A3,A4], in addition to the assumption [A] of 2.3, we will control the FWER at a desired level. First denote by $S(\mathbf{U}, V_1, V_2, V_3, \alpha)$ the admissible set of triples $(\theta_V, \theta_G, \epsilon)$ satisfying:

$$\mu\left(\xi : \bigcup_{v\in V_1}\left(T_v(\xi\odot\mathbf{U}) < \theta_V, T_{g(v)}(\xi\odot\mathbf{U}) < \theta_G\right)\bigcup(\Delta(\xi\odot\mathbf{U}, V_2, V_3, \theta_G) \geq \epsilon)\right) \leq \alpha.$$

Also, keeping the notations of section 2.3, define $V_{00}(\mathbf{U}) := \{v \in V : g(v) \in G_0(\mathbf{U})\}$ and $V_{01}(\mathbf{U}) = V \setminus V_{00}(\mathbf{U})$, where:

$$G_0(\mathbf{U}) = \{g \in G : T_g(\mathbf{U}) > t_0\}.$$

The dependent coarse-to-fine procedure is performed in these two steps:

Algorithm 1 Dependence adapted coarse-to fine procedure

1: Compute $\hat{t}(\mathbf{U}, \alpha) = sup\{t \in I : (\theta_G(t), \theta_V(t), \epsilon(t)) \in S(\mathbf{U}, V, V_{00}(\mathbf{U}), V_{01}(\mathbf{U}), \alpha)\}$
2: Define the rejection set as:

$$\hat{A}(\mathbf{U}, \hat{t}(\mathbf{U}, \alpha)) = \{v \in V : T_v(\mathbf{U}) < \theta_V(\hat{t}(\mathbf{U}, \alpha)) - \epsilon(\hat{t}(\mathbf{U}, \alpha)) \cap T_g(\mathbf{U}) < \theta_G(\hat{t}(\mathbf{U}, \alpha))\}.$$

The dependence adapted coarse-to-fine procedure defined via these two steps con-

trols the FWER at the desired level:

Theorem 3.1.1. *Under the assumptions [A1-4] and [A], the dependent coarse-to-fine procedure defined by* $\hat{A}(\mathbf{U}, \hat{t}(\mathbf{U}, \alpha))$ *controls the FWER at a level* α *:*

$$\mathbb{P}\left(\hat{A}(\mathbf{U}, \hat{t}(\mathbf{U}, \alpha)) \cap A^c \neq \emptyset\right) \leq \alpha.$$

Remark: compared to the previous chapter, assumption [A] is central for the definition of the algorithm. In fact, in Chapter 2, assumption [A] was only necessary for deriving an upper bound of the number of indices in active cells J. If J is given, one can use the Bonferroni coarse-to-fine procedure without assuming [A]. For the dependence adapted coarse-to-fine procedure, we assume [A] so that $V_{00}(\mathbf{U})$ is almost surely included in V_{00}. Therefore, in order to get rid of assumption [A] here, we need to assume that we are given a set that is contained in V_{00}, which is a much stronger assumption than knowing J.

Proof. First, remark that by assumption [*A*], $V_{00}(\mathbf{U}) \subset V_{00}$ and $V_0 \setminus V_{00} \subset V_{01}(\mathbf{U})$. Therefore, we have that:

$$\min_{v \in V_{00}(\mathbf{U}) \cap R(\mathbf{U}, \theta_G)} T_v(\mathbf{U}) - \min_{v \in V_{01}(\mathbf{U})} T_v(\mathbf{U}) \geq \min_{v \in V_{00} \cap R(\mathbf{U}, \theta_G)} T_v(\mathbf{U}) - \min_{v \in V_0 \setminus V_{00}} T_v(\mathbf{U}),$$

and:

$$\Delta(\mathbf{U}, V_{00}(\mathbf{U}), V_{01}(\mathbf{U}), \theta_G) \geq \Delta(\mathbf{U}, V_{00}, V_0 \setminus V_{00}, \theta_G),$$

so that:

$$S(\mathbf{U}, V, V_{00}(\mathbf{U}), V_{01}(\mathbf{U}), \alpha) \supset S(\mathbf{U}, V_{00}, V_{00}, V_0 \setminus V_{00}, \alpha).$$

Hence, if we define $\hat{t}_0(\mathbf{U}, \alpha)$ as:

$$\hat{t}_0(\mathbf{U}, \alpha) := sup\{t \in I : (\theta_G(t), \theta_V(t), \epsilon(t)) \in S(\mathbf{U}, V_{00}, V_{00}, V_0 \setminus V_{00}, \alpha)\},$$

we have that $\hat{t}_0(\mathbf{U}, \alpha) \geq \hat{t}(\mathbf{U}, \alpha)$. Hence $\hat{A}(\mathbf{U}, \hat{t}(\mathbf{U}, \alpha) \subset \hat{A}(\mathbf{U}, \hat{t}_0(\mathbf{U}, \alpha))$. We then rewrite the event $\hat{A}(\mathbf{U}, \hat{t}_0(\mathbf{U}, \alpha)) \cap A^c \neq \emptyset$ as $\bigcup_{v \in A^c}(T_v(\mathbf{U}) < \theta_V(\hat{t}_0(\mathbf{U}, \alpha)) - \epsilon(\hat{t}_0(\mathbf{U}, \alpha)); T_g(\mathbf{U}) < \theta_G(\hat{t}_0(\mathbf{U}, \alpha))$. The intersection of this event with the event $\Delta(\mathbf{U}, V_{00}, V_0, \theta_G(\hat{t}_0(\mathbf{U}, \alpha))) < \epsilon(\hat{t}_0(\mathbf{U}, \alpha))$, is included in the event:

$$\bigcup_{v \in V_{00}} \left(T_v(\mathbf{U}) < \theta_V(\hat{t}_0(\mathbf{U}, \alpha)); T_g(\mathbf{U}) < \theta_G(\hat{t}_0(\mathbf{U}, \alpha)\right),$$

which implies that $\hat{A}(\mathbf{U}, \hat{t}_0(\mathbf{U}, \alpha)) \cap A^c \neq \emptyset$ is included in:

$$\bigcup_{v\in V_{00}} \left(T_v(\mathbf{U}) < \theta_V(\hat{t}_0(\mathbf{U},\alpha)); T_g(\mathbf{U}) < \theta_G(\hat{t}_0(\mathbf{U},\alpha)\right) \bigcup$$

$$\Delta(\xi \odot \mathbf{U}, V_{00}, V_0, \theta_G(\hat{t}_0(\mathbf{U},\alpha))) \geq \epsilon(\hat{t}_0(\mathbf{U},\alpha)).$$

We will therefore prove that the probability of the last event is less than α. Let us define the event $\hat{A}_0(\mathbf{U},t)$ as:

$$\bigcup_{v\in V_{00}} (T_v(\mathbf{U}) < \theta_V(t); T_g(\mathbf{U}) < \theta_G(t)) \bigcup$$

$$\Delta(\mathbf{U}, V_{00}, V_0 \setminus V_{00}, \theta_G(t)) \geq \epsilon(t).$$

We therefore need to prove that $\mathbb{P}\left(\hat{A}_0(\mathbf{U},\hat{t}_0(\mathbf{U},\alpha))\right) \leq \alpha$. At this point let us mention that for every $\xi \in \mathfrak{S}$, we have:

$$S(\mathbf{U}, V_{00}, V_{00}, V_0 \setminus V_{00}, \alpha) = S(\xi \odot \mathbf{U}, V_{00}, V_{00}, V_0 \setminus V_{00}, \alpha).$$

This in turn implies that for every $\xi \in \mathfrak{S}$, $\hat{t}_0(\xi \odot \mathbf{U}, \alpha) = \hat{t}_0(\mathbf{U}, \alpha)$. Using this fact together with the property that the probability of $\hat{A}_0(\mathbf{U}, \alpha, \hat{t}_0(\mathbf{U}, \alpha))$ is invariant under the action of any $\xi \in \mathfrak{S}$, we have

$$\mathbb{P}\left(\hat{A}_0(\mathbf{U}, \hat{t}_0(\mathbf{U}, \alpha))\right) = \mathbb{E}\left(\mu\left(\xi : \hat{A}_0(\xi \odot \mathbf{U}, \hat{t}_0(\mathbf{U}, \alpha))\right)\right).$$

Recall now that by definition of $\hat{t}_0(\mathbf{U}, \alpha)$ and lemma 3.1.1, $\hat{t}_0(\mathbf{U}, \alpha) \in S(\mathbf{U}, V_{00}, V_{00}, V_0 \setminus V_{00}, \alpha)$ and this implies :

$$\mu\left(\xi : \hat{A}_0(\xi \odot \mathbf{U}, \hat{t}_0(\mathbf{U}, \alpha))\right) \leq \alpha,$$

which terminates the proof. □

3.2 Dependence adapted coarse-to-fine procedure based on finite resampling

Here again, the procedure described in 3.1 requires the generation of an infinite number of elements $\xi \in \mathfrak{S}$. In practice, we will be able to generate uniformly a finite number K of elements $\xi_1, \xi_2, ...\xi_K$ which are independent. We will keep the notations of section 2.2, and present three different ways of addressing this issue. The first method that we will present will be exactly the empirical version of the asymptotic resampling procedure, in the sense that we will replace the steps of the former procedure requiring asymptotic resampling by their empirical versions, taking into account the error of estimation coming from the finite resampling. This method will however require a very large K in the sense that the bound on the FWER derived

for this method will require a number K of ξ's generated of order 10^7 if we want the FWER to be controlled at levels belonging to the usual range (less than 0.1). We will then slightly modify the asymptotic resampling based procedure in two different ways and get two procedures that will control the FWER at a desired level with a lower number of permutations K of ξ's $\in \mathfrak{S}$ needed to be generated.

3.2.1 First method

We will now define the set $\hat{S}(\mathbf{U}, \boldsymbol{\xi}, V_1, V_2, V_3, \alpha)$ that will replace the set $S(\mathbf{U}, V_1, V_2, V_3, \alpha)$. It is the set of triples $(\theta_V, \theta_G, \epsilon)$ satisfying:

$$\hat{\mu}\left(\xi : \bigcup_{v \in V_1} \left(T_v(\xi \odot \mathbf{U}) < \theta_V, T_{g(v)}(\xi \odot \mathbf{U}) < \theta_G\right) \bigcup \Delta(\xi \odot \mathbf{U}, V_2, V_3, \theta_G) \geq \epsilon\right) \leq \alpha. \tag{3.1}$$

The dependent coarse-to-fine procedure with finite resampling scores will here be the following:

Algorithm 2 Dependence adapted coarse-to-fine procedure with finite resampling: first method.

1: Compute:

$$\hat{t}^K(\mathbf{U}, \boldsymbol{\xi}, \alpha, \epsilon_1) = sup\{t \in I : (\theta_G(t'), \theta_V(t'), \epsilon(t')) \in \hat{S}(\mathbf{U}, \boldsymbol{\xi}, V, V_{00}(\mathbf{U}), V_{01}(\mathbf{U}), \alpha - \epsilon_1)\}$$

2: Define the rejection set as:

$$\hat{A}(\mathbf{U}, \boldsymbol{\xi}, \hat{t}^K(\mathbf{U}, \boldsymbol{\xi}, \alpha, \epsilon_1)) = \{v \in V : T_v(\mathbf{U}) < \theta_V(\hat{t}^K(\mathbf{U}, \boldsymbol{\xi}, \alpha, \epsilon_1)) - \epsilon(\hat{t}^K(\mathbf{U}, \alpha)) \cap T_{g(v)}(\mathbf{U}) < \theta_G(\hat{t}^K(\mathbf{U}, \boldsymbol{\xi}, \alpha, \epsilon_1))\}.$$

Notice here that the parameter $\hat{t}^K(\mathbf{U}, \boldsymbol{\xi}, \alpha, \epsilon_1)$ depends on the random elements $\xi_1, \xi_2, ..., \xi_K$ that are generated. We now provide an upper bound on the FWER of this procedure:

Theorem 3.2.1. *Under the assumptions [A1-4] and [A], for any positive integer K and $\epsilon > \sqrt{\frac{2}{K}}$:*

$$\mathbb{P}\left(\hat{A}(\mathbf{U}, \boldsymbol{\xi}, \hat{t}^K(\mathbf{U}, \boldsymbol{\xi}, \alpha, \epsilon_1)) \cap A^c \neq \emptyset\right) \leq \alpha + 2K^2 \exp\{2 - \frac{K\epsilon_1^2}{8}\}.$$

Figure 3.1 gives an idea on the number K of generated elements ξ as a function of ϵ for a fixed error term. By error term we mean the term $2K^2 \exp\{2 - \frac{K\epsilon_1^2}{8}\}$. For this we plotted the level curves of the logarithm in base 10 of this term.

Before proving the theorem, let us remind that in Chapter 3, the scores $T_v(\mathbf{U})$ and $T_g(\mathbf{U})$ can be any scores. In particular, they do not require generating an infinite number of ξ's belonging to $\mathfrak{S}$ in order to compute them. To prove theorem 3.2.1, we

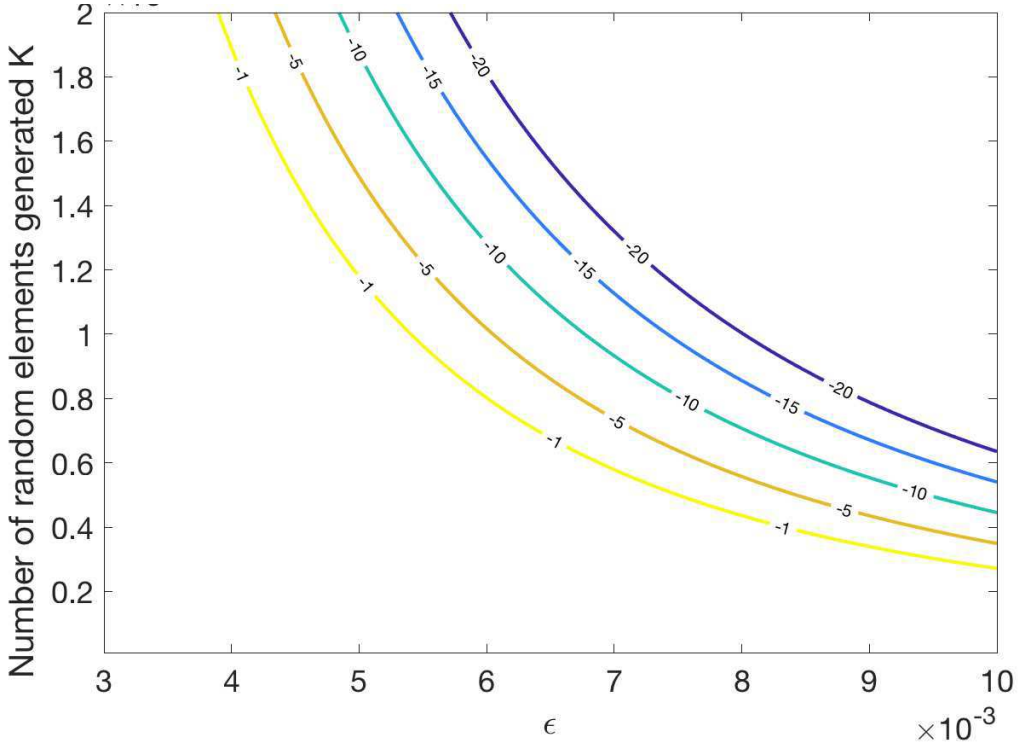

Figure 3.1: Level curves of the logarithm in base 10 of the error term in theorem 3.2.1

will use a version of the Vapnik-Chervonenkis theorem[38–40] that we restate according to our context in the following way:

Theorem 3.2.2 (Vapnik and Chervonenkis)**.** *Let $\mathcal{F}$ be a class of binary functions from $\mathbb{R}^2$ to $\{0,1\}$. Let $s(\mathcal{F}, 2K)$ be the shattering number of order $2K$ of the class $\mathcal{F}$. The shattering number of order k of the class $\mathcal{F}$ for any integer k being defined as:*

$$\max_{(x_1,x_2,....x_k)\in\mathbb{R}^{2\times k}} \{|\,(f(x_1),...f(x_k), f\in\mathcal{F})\,|\}.$$

Then, for every $\epsilon > \sqrt{\frac{2}{K}}$:

$$\hat{\mu}\,(\boldsymbol{\xi} : sup_{f\in\mathcal{F}}((\hat{\mu}-\mu)f > \epsilon)) \leq 2s(\mathcal{F}, 2K)\exp\{-n\frac{K^2}{8}\},$$

where $\hat{\mu}f$ and μf are respectively the expectations of the function f under the empirical measure $\hat{\mu}$ and the asymptotic resampling measure μ, where we interpret points in $\mathbb{R}^2$ as measurable functions of the elements ξ.

We will also combine 3.2.2 with the following lemma :

Lemma 3.2.1. *Under the assumptions [A1, A2, A3, A4], for every subsets V_1, V_2 and V_3 of V, there exist two functions $\phi_1 : \mathcal{U} \to \mathbb{R}$ and $\phi_2 : \mathcal{U} \to \mathbb{R}$ such that for every $t \in I$:*

$$\{\bigcup_{v\in V_1}\big(T_v(\mathbf{U})<\theta_V(t),T_{g(v)}(\mathbf{U})<\theta_G(t)\big)\bigcup\Delta(\mathbf{U},V_2,V_3,\theta_G(t))\geq\epsilon(t)\}$$

is equal to:

$$\{\phi_1(\mathbf{U})<t\cup\phi_2(\mathbf{U})\geq t.\}$$

Proof. To prove lemma 3.2.1, we will construct ϕ_1 and ϕ_2 such that:

$$\bigcup_{v\in V_1}\big(T_v(\mathbf{U})<\theta_V(t),T_{g(v)}(\mathbf{U})<\theta_G(t)\big)\iff\phi_1(\mathbf{U})<t \tag{3.2}$$

and

$$\Delta(\mathbf{U},V_2,V_3,\theta_G(t))\geq\epsilon(t)\iff\phi_2(\mathbf{U})\geq t \tag{3.3}$$

For this, simply define:

$$\phi_1(\mathbf{U}):=sup_{t'\in I}\{\bigcap_{v\in V_1}\big(T_v(\mathbf{U})\geq\theta_V(t')\cup T_{g(v)}(\mathbf{U})\geq\theta_G(t')\big)\}$$

and

$$\phi_2(\mathbf{U}):=sup_{t'\in I}\{\Delta(\mathbf{U},V_2,V_3,\theta_G(t'))\geq\epsilon(t')\}.$$

Indeed, with ϕ_1 and ϕ_2 defined this way, 3.2 and 3.3 hold because, for any $t_1 \leq t_2$ we have:

$$\bigcap_{v \in V_1} \left(T_v(\mathbf{U}) \geq \theta_V(t_2) \cup T_{g(v)}(\mathbf{U}) \geq \theta_G(t_2) \right) \Rightarrow \bigcap_{v \in V_1} \left(T_v(\mathbf{U}) \geq \theta_V(t_1) \cup T_{g(v)}(\mathbf{U}) \geq \theta_G(t_1) \right),$$

and:

$$\Delta(\mathbf{U}, V_2, V_3, \theta_G(t_1)) \geq \epsilon(t_1) \Rightarrow \Delta(\mathbf{U}, V_2, V_3, \theta_G(t_2)) \geq \epsilon(t_2).$$

Moreover, using the same arguments as in lemma 3.1.1, we know that

$$t' \to \mathbf{1}_{\{\bigcap_{v \in V_1} \left(T_v(\mathbf{U}) \geq \theta_V(t') \cup T_{g(v)}(\mathbf{U}) \geq \theta_G(t') \right)\}}$$

and

$$t' \to \mathbf{1}_{\{\Delta(\mathbf{U}, V_2, V_3, \theta_G(t')) \geq \epsilon(t')\}}$$

are left continuous, which terminates the proof of lemma 3.2.1.

□

We are now ready to prove theorem 3.2.1:

Proof of theorem 3.2.1 . Here again, we first notice that the event $\hat{A}(\mathbf{U}, \boldsymbol{\xi}, \hat{t}^K(\mathbf{U}, \boldsymbol{\xi}, \alpha, \epsilon_1)) \cap A^c \neq \emptyset$ is included in the event $\hat{A}(\mathbf{U}, \boldsymbol{\xi}, \hat{t}_0^K(\mathbf{U}, \boldsymbol{\xi}, \alpha, \epsilon_1) \cap A^c \neq \emptyset)$ where $\hat{t}_0^K(\mathbf{U}, \boldsymbol{\xi}, \alpha, \epsilon_1)$

is defined as:

$$\hat{t}_0^K(\mathbf{U}, \boldsymbol{\xi}, , \alpha) := sup\{t \in I : (\theta_G(t), \theta_V(t), \epsilon(t)) \in \hat{S}(\mathbf{U}, \boldsymbol{\xi}, V_{00}, V_{00}, V_0 \setminus V_{00}, \alpha)\}.$$

The inclusion is true because $\hat{t}_0^K(\mathbf{U}, \boldsymbol{\xi}, \alpha - \epsilon_1) \geq \hat{t}^K(\mathbf{U}, \boldsymbol{\xi}, \alpha - \epsilon_1)$. Keeping the notations of the proof of theorem 3.1.1, we will therefore show that:

$$\mathbb{P}\left(\hat{A}_0(\mathbf{U}, \hat{t}_0^K(\mathbf{U}, \boldsymbol{\xi}, \alpha - \epsilon_1))\right) \leq \alpha + 2K^2 \exp\{2 - \frac{K\epsilon_1^2}{8}\}.$$

By theorem 3.1.1, the event $\hat{A}_0(\mathbf{U}, \hat{t}_0^K(\mathbf{U}, \boldsymbol{\xi}, \alpha - \epsilon_1)) \cap \{\hat{t}_0^K(\mathbf{U}, \boldsymbol{\xi}, \alpha - \epsilon_1) \leq \hat{t}_0(\mathbf{U}, \alpha)\}$ is less than α. Therefore it is sufficient to prove that:

$$\mathbb{P}\left(\{\hat{t}_0^K(\mathbf{U}, \boldsymbol{\xi}, \alpha - \epsilon_1) > \hat{t}_0(\mathbf{U}, \alpha)\right) \leq 2K^2 \exp\{2 - \frac{K\epsilon_1^2}{8}\}.$$

Therefore, it is sufficient in turn to prove that $\mathbb{P}\left(\mu\left(\xi : \hat{A}_0(\mathbf{U}, \hat{t}_0^K(\mathbf{U}, \boldsymbol{\xi}, \alpha - \epsilon_1)) > \alpha\right)\right)$ is less than $K^2 \exp\{2 - \frac{K\epsilon_1^2}{8}\}$. In order to prove this result, we will prove that:

$$\mathbb{P}\left(sup_{t \in I}\{\hat{\mu}(\xi : \hat{A}_0(\xi \odot \mathbf{U}, t)) - \mu(\xi : \hat{A}_0(\xi \odot \mathbf{U}, t))\} > \epsilon_1\right) \leq 2K^2 \exp\{2 - \frac{K\epsilon_1^2}{8}\}.$$

Using lemma 3.2.1, we know that there exist two functions ϕ_1 and ϕ_2 such that $\mathbf{1}_{\hat{A}_0(\xi \odot \mathbf{U}, t)}$ is nothing but $\mathbf{1}_{\{\phi_1(\xi \odot \mathbf{U}) < t \cup \phi_2(\xi \odot \mathbf{U}) \geq t\}}$. Therefore, using theorem 3.2.2:

$$\mathbb{P}\left(sup_{t\in I}\{\hat{\mu}(\xi : \hat{A}_0(\xi \odot \mathbf{U}, t)) - \mu(\xi : \hat{A}_0(\xi \odot \mathbf{U}, t))\} > \epsilon_1\right) \leq 2s(\mathcal{F}, 2K)\exp\{-\frac{K\epsilon_1^2}{8}\},$$

where $\mathcal{F}$ is the class of indicators on the complements of the lower left quadrants on the 2 dimensional plane. The VC dimension of this class is equal to 2. To see this, it suffices to remark that for any configuration of three points in the plane, there will be two points such that it is impossible to include them in lower left quadrants without including the remaining point. therefore, using Sauer's lemma (see for example[40]), $s(\mathcal{F}, 2K) \leq (Ke)^2$, which gives:

$$\mathbb{P}\left(sup_{t\in I}\{\hat{\mu}(\xi : \hat{A}_0(\xi \odot \mathbf{U}, t)) - \mu(\xi : \hat{A}_0(\xi \odot \mathbf{U}, t))\} > \epsilon_1\right) \leq 2K^2\exp\{2 - \frac{K\epsilon_1^2}{8}\},$$

which terminates the proof.

□

3.2.2 Second method

Here, we keep the same definition as in equation 3.1 for the set $\hat{S}(\mathbf{U}, \boldsymbol{\xi}, V_1, V_2, V_3, \alpha)$ that will replace the set $S(\mathbf{U}, V_1, V_2, V_3, \alpha)$.

The dependent coarse-to-fine, second method procedure with finite resampling

scores is the following:

Algorithm 3 Dependence adapted coarse-to-fine procedure with finite resampling: second method.

1: Compute:

$$\hat{t}^K(\mathbf{U}, \boldsymbol{\xi}, \alpha) = sup\{t \in I : (\theta_G(t'), \theta_V(t'), \epsilon(t')) \in \hat{S}(\mathbf{U}, \boldsymbol{\xi}, V, V_{00}(\mathbf{U}), V_{01}(\mathbf{U}), \alpha), \ \forall t' \leq t\}$$

2: Define the rejection set as:

$$\hat{A}(\mathbf{U}, \boldsymbol{\xi}, \hat{t}^K(\mathbf{U}, \boldsymbol{\xi}, \alpha)) = \{v \in V : T_v(\mathbf{U}) < \theta_V(\hat{t}^K(\mathbf{U}, \boldsymbol{\xi}, \alpha)) - \epsilon(\hat{t}^K(\mathbf{U}, \boldsymbol{\xi}, \alpha)) \cap T_{g(v)}(\mathbf{U}) < \theta_G(\hat{t}^K(\mathbf{U}, \boldsymbol{\xi}, \alpha))\}.$$

Notice that as opposed to the infinite resampling scores procedure, here we require that $(\theta_G(t'), \theta_V(t'), \epsilon(t')) \in \hat{S}(\mathbf{U}, V, V_{00}(\mathbf{U}), V_{01}(\mathbf{U}), \alpha) \ \ \forall t' \leq t$ which is more conservative than the condition $(\theta_G(t), \theta_V(t), \epsilon(t)) \in \hat{S}(\mathbf{U}, V, V_{00}(\mathbf{U}), V_{01}(\mathbf{U}), \alpha)$. To see why it is the case, remark that the set $\Delta(\xi \odot \mathbf{U}, V_{00}, V_0 \setminus V_{00}, \theta_G(t)) \geq \epsilon(t)$ shrinks when t increases. More precisely, $\Delta(\xi \odot \mathbf{U}, V_{00}, V_0 \setminus V_{00}, \theta_G(t)) \geq \epsilon(t)$ implies that $\Delta(\xi \odot \mathbf{U}, V_{00}, V_0 \setminus V_{00}, \theta_G(t')) \geq \epsilon(t')$ for every ξ when $t > t'$. This is the case because $\epsilon(t)$ and $\theta_G(t)$ are nondecreasing functions of t. Therefore, we can have situations where $t > t'$ with $(\theta_G(t'), \theta_V(t'), \epsilon(t')) \notin \hat{S}(\mathbf{U}, \boldsymbol{\xi}, V, V_{00}(\mathbf{U}), V_{01}(\mathbf{U}), \alpha)$ but $(\theta_G(t), \theta_V(t), \epsilon(t)) \in \hat{S}(\mathbf{U}, \boldsymbol{\xi}, V, V_{00}(\mathbf{U}), V_{01}(\mathbf{U}), \alpha)$.

Now, we claim that the described procedure controls the FWER at the following desired level.

Theorem 3.2.3. *Under the assumptions [A1-4] and [A], for any positive integer K:*

$$\mathbb{P}\left(\hat{A}(\mathbf{U}, \boldsymbol{\xi}, \hat{t}^K(\mathbf{U}, \boldsymbol{\xi}, \alpha)) \cap A^c \neq \emptyset\right) \leq \frac{\lfloor K\alpha \rfloor + 1}{K + 1}.$$

Proof. First, as it was noticed in the proof for the asymptotic resampling scores case, we have that:

$$\hat{S}(\mathbf{U}, V, V_{00}(\mathbf{U}), V_{01}(\mathbf{U}), \alpha) \subset \hat{S}(\mathbf{U}, V_{00}, V_{00}, V_0, \alpha),$$

and if we define $\hat{t}_0^K(\mathbf{U}, \alpha)$ as:

$$\hat{t}_0^K(\mathbf{U}, \alpha) := sup\{t \in I : (\theta_G(t), \theta_V(t), \epsilon(t)) \in \hat{S}(\mathbf{U}, V_{00}, V_{00}, V_0 \setminus V_{00}, \alpha) \ \ \forall t' \leq t\},$$

we have that $\hat{t}_0^K(\mathbf{U}, \boldsymbol{\xi}, \alpha) \geq \hat{t}^K(\mathbf{U}, \boldsymbol{\xi}, \alpha)$ and:

$$\mathbb{P}\left(\hat{A}(\mathbf{U}, \boldsymbol{\xi}, \hat{t}^K(\mathbf{U}, \boldsymbol{\xi}, \alpha)) \cap A^c \neq \emptyset\right) \leq \mathbb{P}\left(\hat{A}(\mathbf{U}, \boldsymbol{\xi}, \hat{t}_0^K(\mathbf{U}, \boldsymbol{\xi}, \alpha)) \cap A^c\right).$$

Let us define:

$$t^*(\mathbf{U}) = sup\{t \in I : \forall v \in V_{00} : (T_v(\mathbf{U}) \geq \theta_V(t') \ \ or \ \ T_g(\mathbf{U}) \geq \theta_G(t'))$$
$$and \ \ \Delta(\mathbf{U}, V_{00}, V_0 \setminus V_{00}, \theta_G(t')) < \epsilon(t'), \ \ \forall t' \leq t\}.$$

The distribution of $t^*(\mathbf{U})$ is by construction invariant under the action of any $\xi \in \mathfrak{S}$, and:

$$\begin{aligned}
\mathbb{P}\left(\hat{A}(\mathbf{U},\boldsymbol{\xi},\hat{t}_0^K(\mathbf{U},\boldsymbol{\xi},\alpha))\cap A^c\right) &\leq \mathbb{P}\left(\sum_{k=1}^{K}\mathbf{1}_{t^*(\mathbf{U})\leq t^*(\xi_k\odot\mathbf{U})}\leq \lfloor K\alpha\rfloor\right)\\
&\leq \mathbb{E}\left(G_\beta(1-\mu\left(\xi : t^*(\mathbf{U})\leq t^*(\xi\odot\mathbf{U})\right), K-\lfloor K\alpha\rfloor, \lfloor K\alpha\rfloor+1)\right)\\
&\leq \int_0^1 G_\beta(t, K-\lfloor K\alpha\rfloor, \lfloor K\alpha\rfloor+1)dt\\
&\leq \frac{\lfloor K\alpha\rfloor+1}{K+1}
\end{aligned}$$

where the third inequality follows from the fact that $\mu\left(\xi : t^*(\mathbf{U})\leq t^*(\xi\odot\mathbf{U})\right)$ dominates the uniform distribution on $[0,1]$ and $G_\beta(t, K-\lfloor K\alpha\rfloor, \lfloor K\alpha\rfloor+1)$ is increasing in t, which terminates the proof.

□

3.2.3 Dependence adapted coarse-to-fine procedure with finite resampling: third method.

.

For the third method, we will modify the definition of $\hat{S}(\mathbf{U},\boldsymbol{\xi},V_1,V_2,V_3,\alpha)$, in the sense that it will no longer be the natural empirical version of asymptotic resampling set $S(\mathbf{U},V_1,V_2,V_3,\alpha)$. More precisely, for this method, the set $\hat{S}(\mathbf{U},\boldsymbol{\xi},V_1,V_2,V_3,\alpha)$ of triples $(\theta_V,\theta_G,\epsilon)$ will be defined as the triples satisfying:

$$\hat{\mu}\left(\xi : \bigcup_{v\in V_1}\left(T_v(\xi\odot\mathbf{U}) < \theta_V, T_{g(v)}(\xi\odot\mathbf{U}) < \theta_G\right)\right) \leq \frac{\alpha}{2}$$

$$\text{and } \hat{\mu}\left(\xi : \Delta(\xi\odot\mathbf{U}, V_2, V_3, \theta_G) \geq \epsilon\right) \leq \frac{\alpha}{2}.$$

With this definition, the coarse-to-fine method is the following:

Algorithm 4 Finite resampling algorithm: third method

1: Compute:

$$\hat{t}^K(\mathbf{U}, \boldsymbol{\xi}, \alpha) = sup\{t \in I : (\theta_G(t), \theta_V(t), \epsilon(t)) \in \hat{S}(\mathbf{U}, \boldsymbol{\xi}, V, V_{00}(\mathbf{U}), V_{01}(\mathbf{U}), \alpha)\}.$$

2: Define the rejection set as:

$$\hat{A}(\mathbf{U}, \boldsymbol{\xi}, \hat{t}^K(\mathbf{U}, \boldsymbol{\xi}, \alpha)) = \{v \in V : T_v(\mathbf{U}) < \theta_V(\hat{t}^K(\mathbf{U}, \boldsymbol{\xi}, \alpha)) - \epsilon(\hat{t}^K(\mathbf{U}, \boldsymbol{\xi}, \alpha)) \cap T_g(\mathbf{U}) < \theta_G(\hat{t}^K(\mathbf{U}, \boldsymbol{\xi}, \alpha))$$

For this third method, the FWER will be controlled at the following level for any number K of generated elements in $\mathfrak{S}$:

Theorem 3.2.4. *Under the assumptions [A1-4] and [A], for any positive integer K:*

$$\mathbb{P}\left(\hat{A}(\mathbf{U}, \boldsymbol{\xi}, \hat{t}^K(\mathbf{U}, \boldsymbol{\xi}, \alpha)) \cap A^c \neq \emptyset\right) \leq 2\frac{\lfloor \frac{K}{2}\alpha \rfloor + 1}{K+1}..$$

Proof. The starting point of the proof of theorem 3.2.4 is identical to the proof in the other methods:

$$\hat{S}(\mathbf{U}, \boldsymbol{\xi}, V, V_{00}(\mathbf{U}), V_{01}(\mathbf{U}), \alpha) \subset \hat{S}(\mathbf{U}, \boldsymbol{\xi}, V_{00}, V_{00}, V_0 \setminus V_{00}, \alpha),$$

and take $\hat{t}_0^K(\mathbf{U}, \boldsymbol{\xi}, \alpha)$ as:

$$\hat{t}_0^K(\mathbf{U}, \boldsymbol{\xi}, \alpha) := sup\{t \in I : (\theta_G(t), \theta_V(t), \epsilon(t)) \in \hat{S}(\mathbf{U}, \boldsymbol{\xi}, V_{00}, V_{00}, V_0, \alpha)\},$$

we have that $\hat{t}_0^K(\mathbf{U}, \boldsymbol{\xi}, \alpha) \geq \hat{t}^K(\mathbf{U}, \boldsymbol{\xi}, \alpha)$, and we will prove that:

$$\mathbb{P}\left(\hat{A}(\mathbf{U}, \hat{t}_0^K(\mathbf{U}, \boldsymbol{\xi}, \alpha)\right) \leq \alpha.$$

At this point, define:

$$t_1^*(\mathbf{U}) = sup\{t \in I : \forall v \in V_{00} : (T_v(\xi \odot \mathbf{U}) \geq \theta_V(t') \;\; or \;\; T_g(\xi \odot \mathbf{U}) \geq \theta_G(t'))\},$$

and:

$$t_2^*(\mathbf{U}) = sup\{t \in I : \Delta(\xi \odot \mathbf{U}, V_{00}, V_0 \setminus V_{00}, \theta_G(t')) \geq \epsilon(t')\}.$$

Using these two statistics, we bound the probability:

$$\begin{aligned}\mathbb{P}\left(\hat{A}_0(\mathbf{U},\hat{t}_0^K(\mathbf{U},\boldsymbol{\xi},\alpha)\right) &\leq \mathbb{P}\left(\hat{\mu}\left(\xi : t_1^*(\mathbf{U}) \leq t_1^*(\xi\odot\mathbf{U})\right) \leq \frac{\alpha}{2}\bigcup \hat{\mu}\left(t_2^*(\xi\odot\mathbf{U}) \leq t_2^*(\mathbf{U})\right)\right) \leq \frac{\alpha}{2}\Big) \\ &\leq \mathbb{P}\left(\hat{\mu}\left(\xi : t_1^*(\mathbf{U}) \leq t_1^*(\xi\odot\mathbf{U})\right) \leq \frac{\alpha}{2}\right) + \mathbb{P}\left(\hat{\mu}\left(\xi : t_2^*(\xi\odot\mathbf{U}) \leq t_2^*(\mathbf{U})\right) \leq \frac{\alpha}{2}\right. \\ &\leq 2\frac{\lfloor\frac{K}{2}\alpha\rfloor+1}{K+1}.\end{aligned}$$

which terminates the proof.

□

3.3 Simulations

In the first set of simulations, we consider the following toy model. We observe a random variable Y and a set of random variables $(X_v, v \in V)$ such that:

$$Y = \sum_{v\in A} X_v + 5\eta,$$

where η is a standard Gaussian independent from the variables $(X_v, v \in V)$. The set A has size 25 and all the variables $(X_v, v \in A)$ are i.i.d standard Gaussian random variables independent from $(X_v, v \in A^c)$. As for the random variables $(X_v, v \in A^c)$, they are constructed in the following way. Let $(\tilde{X}_v, v \in A^c)$ be i.i.d standard Gaussian variables and $\tilde{\eta}$ a standard Gaussian variable independent from all the other variables. For each v, $X_v = \tilde{\eta} + \sigma\tilde{X}_v$, where σ is a varying parameter driving

the dependence between the variables $(X_v, v \in V)$. In our simulations, we look at 4 scenarios where σ takes four possibles values: $0.01, 0.1, 1$ and 10. For all of the simulations, the sample size is 600 and the size of each cell is 25. The number of tests $|V|$ is 10^4 . We generate the clustering according to the the Chinese restaurant process described in section 2.5 with a clustering parameter that is equal to 1. We will compare the performance of the second dependence adapted coarse-to-fine method, the third dependence adapted coarse-to-fine method, the coarse-to-fine Bonferroni, the single level dependence adapted method that is described in section 1.6 and finally the Bonferroni-Holm method with the default thresholds setting. For the dependence adapted coarse-to-fine procedures, we took $\theta_V(t) = \theta_G(t) = t$ and $\epsilon(t) = \frac{t}{10}$ with $t \in [0, 1]$. The number of permutations performed for each simulation is 10^5 and for each σ, we average the performance of 50 independent simulations. The FWER is controlled at 0.1. Finally, let us mention that we did not include the first dependence adapted coarse-to-fine procedure because it turned out to be identical to the second dependence adapted coarse-to-fine procedure for these simulations. The results obtained are summarized in table 3.1. The interpretation of the last column is simple, the performance of the Bonferroni-Holm estimator does not depend on the parameter σ. In fact, we averaged over all the simulations for the column corresponding to the Bonferroni-Holm method. The interpretation of the column corresponding to the dependence adapted one level method is also simple: the smaller is sigma, the higher is the dependence between the variables $(T_v(\mathbf{U}), v \in A^c)$ and the

larger will be the estimated quantiles of the statistic $\min_{v\in A^c} T_v(\mathbf{U})$. As a consequence, the number of detections is higher. As for the Bonferroni coarse-to-fine method, as mentioned earlier in this section, the quality of the estimator $\hat{J}$ deteriorates when the dependence between the tests is higher, and it is clear that the number of detections of this method increases with the parameter σ. Finally, for the two remaining methods, the situation is more complicated. First, notice that they outperform the remaining methods for all values of σ (Recall that the clustering parameter is fixed here). The second dependence adapted coarse-to-fine method is by definition less conservative than the third method, and this is observed in the two first columns of table 3.1. The more surprising effect for these two columns is that the performance is not decreasing with the value of σ. The reason is that the effect σ on these two methods is a mix between the effect on the Bonferroni coarse-to-fine method and the dependence adapted one level method. More precisely, using the notations of lemma 3.2.1 with $V_1 = V, V_2 = V_{00}(\mathbf{U})$ and $V_3 = V_{01}(\mathbf{U})$, the smaller is σ, the larger the estimated quantiles of the function $\phi_1(\mathbf{U})$ will be, which leads to less conservative thresholds and more detections. The situation for the function ϕ_2 is the opposite. The smaller is the parameter σ, the larger the estimated quantiles of the function $\phi_2(\mathbf{U})$ will be , which leads to more conservative thresholds and less detections. Notice however that the two procedures seem to take advantage of the dependence when the dependence is not extreme.

In the second set of simulations, we used PLINK generated datasets and the setting

Value of σ	Dep adapted CTF 2	Dep adapted CTF 3	Bonferroni CTF	Dep adapted	BH
0.01	16.12	13.48	5.54	9.18	5.36
0.1	18.46	17.22	7.72	8.66	5.36
1	18.28	17.56	8.9	8.14	5.36
10	16.72	16.54	12.16	6.2	5.36

Table 3.1: Average number of true detections for each of the 5 methods, from left to right: second dependence adapted coarse-to-fine method, third dependence adapted coarse-to-fine method , Bonferroni coarse-to-fine, dependence adapted one level method and Bonferroni-Holm.

is identical to Section 2.5. We generated datasets with parameters from section 2.5, with 10^5 permutations for each simulation, iterating over $\alpha = 0.5$, 1, 5, 10, 20, 30, 40 and 50 in the Chinese restaurant process and considering four cases: (I) $\nu_G = 10$, no LD; (II) $\nu_G = 25$, no LD; (III) $\nu_G = 10$, LD = 0.8; (III) $\nu_G = 25$, LD = 0.8. In each case, we took the average over 50 simulations. The only difference is that the active set has size 50 instead of 25.

Tables 3.2, 3.3, 3.4 and 3.5 summarize the results obtained under the different settings.

Table 3.2: Cell size 10, independent SNPs

Value of α	Dep adapted CTF 2	Dep adapted CTF 3	Bonferroni CTF	Dep adapted	BH
0.5	19.68	19.46	17.6	13.62	13.62
1	19.06	18.86	16.12	13.62	13.62
5	17.48	17.14	14.58	13.62	13.62
10	16.22	15.96	13.7	13.62	13.62
20	15.04	14.62	12.02	13.62	13.62
30	14.08	13.18	11.16	13.62	13.62
40	12.46	11.94	10.34	13.62	13.62
50	11.58	10.66	10.08	13.62	13.62

Table 3.3: Cell size 10, correlated SNPs

Value of α	Dep adapted CTF 2	Dep adapted CTF 3	Bonferroni CTF	Dep adapted	BH
0.5	22.74	20	12.72	9.27	8.73
1	22.1	19.42	12..34	9.27	8.73
5	20.32	17.68	10.62	9.27	8.73
10	19.04	15.94	10.46	9.27	8.73
20	17.66	14.26	9.24	9.27	8.73
30	15.58	13.18	8.9	9.27	8.73
40	14.46	12.3	6.58	9.27	8.73
50	11.12	10.36	6.32	9.27	8.73

Table 3.4: Cell size 25, independent SNPs

Value of α	Dep adapted CTF 2	Dep adapted CTF 3	Bonferroni CTF	Dep adapted	BH
0.5	13.26	11.98	7.74	4.29	4.27
1	12.24	10.72	7.72	4.29	4.27
5	9.96	8.5	5.58	4.29	4.27
10	8.1	6.82	3.92	4.29	4.27
20	6	5.24	3.54	4.29	4.27
30	4.78	4.1	2.82	4.29	4.27
40	4.06	3.76	2.78	4.29	4.27
50	3.62	3.42	2.66	4.29	4.27

Table 3.5: Cell size 25, correlated SNPs

Value of α	Dep adapted CTF 2	Dep adapted CTF 3	Bonferroni CTF	Dep adapted	BH
0.5	24.48	20.94	14.28	10.23	9.96
1	23.16	20.46	12.92	10.23	9.96
5	20.54	17.86	10.46	10.23	9.96
10	18.24	15.64	9.3	10.23	9.96
20	15.84	13.08	7.44	10.23	9.96
30	13.76	11.42	5.32	10.23	9.96
40	12.02	9.74	4.56	10.23	9.96
50	10.9	10.12	4.38	10.23	9.96

Chapter 4

Coarse-to-fine procedures controlling the FDR

As presented in the introductory part, the other popular criterion for controlling errors in multiple hypotheses testing is the FDR. Several procedures for controlling the FDR have been proposed, and we described the main contributions in Section 1.3. Contrarily to procedures aiming at controlling the FWER, the procedures controlling the FDR usually require additional assumptions. Two types of assumptions are usually made: either the tests are independent, or we assume positive regression dependence on each of a subset (PRDS).[8] This property will be precisely defined in the next section. In the coarse-to-fine setting, the assumption of independence between the tests does not make sense. The reason is that for each v, the decision of the rejection or not of hypotheses $H_0(v)$ depends on the statistic $T_{g(v)}(\mathbf{U})$. Evidently,

unless v is the unique index belonging to the cell $g(v)$, the tests associated to two indices in the same cell will not be independent, even when the statistics $T_v(\mathbf{U})$ are independent. We will therefore work under the PRDS framework. In fact, we will derive a coarse-to-fine procedure controlling the FDR under that assumption, in addition to the assumption that J is a known integer. We will see that unlike the FWER setting, we will need an additional assumption if we want to replace J by an estimated upper bound $\hat{J}$. There exists a less popular family of procedures summarized by,[10] controlling the FDR under any dependence structure. However, these procedures are very conservative and usually comparable to Bonferoni procedures. This leads to a number of detections significantly lower than the widely used procedures for controlling the FDR.

We propose coarse-to-fine procedures versions of this family of procedures, in order to improve the power of detection when the clustering assumption is true. As a consequence, we will illustrate via simulations how we can get algorithms controlling the FDR under any dependence setting, and having a power of detection comparable to the Benjamini-Hochbebrg procedure, provided that the clustering assumption is true. The chapter is organized as follows: we first propose a procedure that controls the FDR under the PRDS assumption and assumes that J is a known integer. We then present a procedure replacing J by an estimated upper bound $\hat{J}$ and show that we control the FDR provided that we add an assumption slightly stronger than the PRDS assumption. Finally, we derive a family of procedures controlling the FDR

under any dependence assumption between the tests. Let us mention that all the procedures will be presented in the asymptotic resampling framework. The reason is that the modifications required for the finite resampling case is immediate. We will briefly describe how it is done. Finally, we compare our coarse-to-fine procedures with the Benjamini-Hocheberg procedure and examples from the family of procedures controlling the FDR under any dependence between tests that described in.[10] The comparison is done via simulation using the PLINK dataset under the same setting as the previous chapters.

4.1 Asymptotic resampling procedure

In this part, we propose a coarse-to-fine method designed to control the False Discovery Rate. The procedure described in this part will correspond to choosing the thresholds at the cell level and at the indices level adaptively (with respect to the data). Recall that we denote the scores at the index level for any $v \in V$ by $\rho_v(\mathbf{U})$, where ρ_v is a real valued function from $\mathcal{U}$ to $\mathbb{R}$. The scores at the cell level, for any cell $g \in G$ will be denoted by $\rho_g(\mathbf{U})$ where ρ_g is a real valued function from $\mathcal{U}$ to $\mathbb{R}$. We define two additional functions ϕ_v and ϕ_g that we assume to be nondecreasing functions from $\mathbb{R}$ to $\mathbb{R}$. Also, we define two asymptotic resampling scores for every $v \in V$ as following:

$$T_v^0(\mathbf{U}) := \frac{1}{|\mathfrak{S}|}\sum_{\xi\in\mathfrak{S}} \mathbf{1}_{\rho_v(\mathbf{U})\leq\rho_v(\xi\odot\mathbf{U})},$$

and

$$T_v^{00}(\mathbf{U}) := \frac{1}{|\mathfrak{S}|}\sum_{\xi\in\mathfrak{S}} \mathbf{1}_{\min(\phi_V(\rho_v(\mathbf{U})),\phi_G(\rho_{g(v)}(\mathbf{U})))\leq\min(\phi_V(\rho_v(\xi\odot\mathbf{U})),\phi_G(\rho_{g(v)}(\xi\odot\mathbf{U})))}.$$

4.1.1 Procedure with known J assumption

The detection set will be a function of $(T_v^0(\mathbf{U}), T_v^{00}(\mathbf{U}))_{v\in V}$. Before describing the procedure, let us first state the assumptions under which assumptions our procedure will control the FDR at a desired level:

- ⊙ A1. We first assume that $J = |V_0 \setminus V_{00}|$ is known. We will later relax this assumption at the cost of strengthening the assumption that will follow.
- ⊙ A2. We also assume that the random vector $(T_v^0(\mathbf{U}), T_v^{00}(\mathbf{U}))_{v\in V}$ is PRDS (positive regression dependence on each of a subset) on the vector $(T_v^0(\mathbf{U}), T_v^{00}(\mathbf{U}))_{v\in V}$. We define below the notion of positive regression dependency.

Definition 4.1.1. *We define the PRDS (positive regression dependence on each of a subset) property as following:*

- ⊙ *For any integer k, we say that $x \geq y$ for any two vectors x and y in $\mathbb{R}^k$ if and*

only if $x_i \geq y_i$ *for every* $i \in [1:k]$.

- ⊙ *For any integer* k *and any set* $D \subset \mathbb{R}^k$*, we say that* D *is a nondecreasing set if and only if* $y \in D$ *and* $x \geq y$ *implies that* $x \in D$*. Finally, for any* $\mathbb{R}^k$ *valued random vector* $X = (X_i)_{i\in[1:k]}$*, and any random sub-vector of* X *denoted by* $Y = (X_i)_{i\in I}$ *and* $I \subset \{1, 2, ..k\}$*, we will say that* X *is PRD on each of* Y *if for every* $i \in I$*, the function defined as:*

$$x_i \to \mathbb{P}\left(X \in D | X_i = x_i\right),$$

is a nondecreasing function for any nondecreasing set $D \subset \mathbb{R}^k$.

The definition of the PRDS notion of a vector X on a sub-vector Y implies a property that will play a central role in controlling the FDR in our procedure.

Lemma 4.1.1. *let* k *be any integer and an* $\mathbb{R}^k$ *valued random vector* X*. Let* $Y = (X_i)_{i\in I}$ *and* $I \subset \{1, 2, ..k\}$ *be any random sub-vector of* X*. if* X *is PRDS on* Y*, then for any* $i \in I$ *and nondecreasing set* $D \subset \mathbb{R}^k$*:*

$$x_i \to \mathbb{P}\left(X \in D | X_i \leq x_i\right),$$

is a nondecreasing function of x_i.

Proof. let $t_1 \leq t_2$ be any two real numbers. We have

$$\begin{aligned}
\mathbb{P}\left(X \in D | X_i \leq t_2\right) &= \mathbb{E}\left(\mathbb{P}\left(X \in D | X_i\right) | X_i \leq t_2\right) \\
&= \mathbb{E}\left(\mathbb{P}\left(X \in D | X_i\right) \mathbf{1}_{t_1 < X_i \leq t_2} | X_i \leq t_2\right) + \mathbb{E}\left(\mathbb{P}\left(X \in D | X_i\right) \mathbf{1}_{X_i \leq t_1} | X_i \leq t_2\right) \\
&\geq \mathbb{E}\left(\mathbb{P}\left(X \in D | X_i \leq t_1\right) \mathbf{1}_{t_1 < X_i \leq t_2} | X_i \leq t_2\right) + \mathbb{E}\left(\mathbb{P}\left(X \in D | X_i\right) \mathbf{1}_{X_i \leq t_1} | X_i \leq t_2\right. \\
&= \mathbb{P}\left(X \in D | X_i \leq t_1\right) \mathbb{P}\left(t_1 < X_i \leq t_2 | X_i \leq t_2\right) \\
&\quad + \mathbb{P}\left(X \in D | X_i \leq t_1\right) \mathbb{P}\left(X_i \leq t_1 | X_i \leq t_2\right)
\end{aligned}$$

We deduce the third from the second line because:

$$\mathbb{P}\left(X \in D | X_i\right) \mathbf{1}_{t_1 < X_i \leq t_2} \geq \mathbb{P}\left(X \in D | X_i = t_1\right) \mathbf{1}_{t_1 < X_i \leq t_2},$$

and

$$\mathbb{P}\left(X \in D | X_i \leq t_1\right) \mathbf{1}_{t_1 < X_i \leq t_2} \leq \mathbb{P}\left(X \in D | X_i = t_1\right) \mathbf{1}_{t_1 < X_i \leq t_2},$$

which terminates the proof. □

We are now ready to describe our coarse-to-fine FDR procedure and the theorem stating that it controls the FDR at a desired level α. First, define for any integer $i \leq |V|$, the family of random subsets of V:

$$L_{00}(\mathbf{U}, i, \alpha) := \{v \in V : T_v^{00}(\mathbf{U}) \leq \frac{\alpha i}{|V|}\},$$

and

$$L_0(\mathbf{U}, i, \alpha, J) := \{v \in V : T_v^{0}(\mathbf{U}) \leq \frac{\alpha i}{J}\}.$$

Using this family of sets, the coarse-to-fine FDR procedure, which we will refer to as the Benjamini-Hochberg coarse-to-fine method, is the following:

Algorithm 5 Benjamini-Hochberg coarse-to-fine with knwon J

1: Compute $i_* = \max\{i \leq |V| : i \leq |L_{00}(\mathbf{U}, i, \alpha) \cap L_0(\mathbf{U}, i, \alpha, J)|\}$
2: Define the rejection set as:

$$\hat{A}(\mathbf{U}, \alpha, J) = L_{00}(\mathbf{U}, i_*, \alpha) \cap L_0(\mathbf{U}, i_*, \alpha, J).$$

Algorithm 5 allows one to control the coarse to fine procedure at a desired level α. More precisely:

Theorem 4.1.1. *Under the assumption [A2.], we have:*

$$\mathbb{E}\left(\frac{|\hat{A}(\mathbf{U}, \alpha, J) \cap A^c|}{|\hat{A}(\mathbf{U}, \alpha, J) \vee 1|}\right) \leq \alpha.$$

Before proving theorem 4.1.1, let us justify that the procedure described by the algorithm 5 is a coarse-to-fine procedure. To see this let us assume, as a simplification, that :

⊙ S1. For every $v \in V$ and $\mathbf{U} \in \mathcal{U}$, we have :

$$F_v(\mathbf{U},.) : t \to \mu\left(\xi : \phi_V(\rho_v(\xi \odot \mathbf{U})), \phi_G(\rho_{g(v)}(\xi \odot \mathbf{U})) \leq t\right) \text{ is continuous.}$$

Under the simplification S1, one can state the following proposition:

Proposition 4.1.1. *Assume that the function ϕ_G and ϕ_V are cadlag (right continuous left limits). Denote by:*

$$\phi_V^*(t) = \inf\{x : \phi_V^-(x) \geq t\},$$

$$\phi_G^*(t) = \inf\{x : \phi_G^-(x) \geq t\}$$

and

$$F_v^*(\mathbf{U}, t) = \inf\{x : F_v^-(\mathbf{U}, x) \geq t\},$$

where ϕ_V^-, ϕ_G^- and F_v^- are respectively the right limits of ϕ_V, ϕ_G and F_v. With these notations, we have, for every $i \leq |V|$:

$$L_{00}(\mathbf{U}, i, \alpha) = \{v \in V : \rho_{g(v)}(\mathbf{U}) \geq \phi_G^*(F_v^*(\mathbf{U}, 1-\frac{i\alpha}{|V|}))\} \text{ and } \rho_v(\mathbf{U}) \geq \phi_V^*(F_v^*(\mathbf{U}, 1-\frac{i\alpha}{|V|}))\}.$$

Proof.

$$
\begin{aligned}
L_{00}(\mathbf{U}, i, \alpha) &= \{v \in V : T_v^{00}(\mathbf{U}) \leq \frac{\alpha i}{|V|}\} \\
&= \{v \in V : 1 - F_v\left(\mathbf{U}, \min(\phi_V(\mathbf{U}), \phi_G(\rho_{g(v)}(\mathbf{U})))\right) \leq \frac{\alpha i}{|V|}\} \\
&= \{v \in V : \min(\phi_V(\rho_v(\mathbf{U})), \phi_G(\rho_{g(v)}(\mathbf{U}))) \geq F_v^*(\mathbf{U}, 1 - \frac{\alpha i}{|V|})\} \\
&= \{v \in V : \rho_v(\mathbf{U}) \geq \phi_V^*(F_v^*(\mathbf{U}, 1 - \frac{\alpha i}{|V|})) \;\; and \;\; \rho_{g(v)}(\mathbf{U}) \geq \phi_G^*(F_v^*(\mathbf{U}, 1 - \frac{\alpha i}{|V|})) \;\; \}
\end{aligned}
$$

□

We now prove theorem 4.1.1:

Proof of theorem 4.1.1. We first notice that the sets $L_{00}(\mathbf{U}, i, \alpha)$ and $L_0(\mathbf{U}, i, \alpha, J)$ are increasing with i, and so are the intersections. Therefore:

$$i \leq |L_{00}(\mathbf{U}, i, \alpha) \cap L_0(\mathbf{U}, i, \alpha, J)|$$

is equivalent to:

$$
\begin{aligned}
&L_{00}\left(\mathbf{U}, i, \alpha\right) \cap L_0\left(\mathbf{U}, i, \alpha, J\right) \subset \\
&L_{00}\left(\mathbf{U}, |L_{00}(\mathbf{U}, i, \alpha) \cap L_0(\mathbf{U}, i, \alpha, J)|, \alpha\right) \cap L_0\left(\mathbf{U}, |L_{00}(\mathbf{U}, i, \alpha) \cap L_0(\mathbf{U}, i, \alpha, J)|, \alpha, J\right).
\end{aligned}
$$

This implies that $\hat{A}(\mathbf{U}, \alpha, J) = L_{00}(\mathbf{U}, i_*, \alpha) \cap L_0(\mathbf{U}, i_*, \alpha, J)$ is included in the set

$L_{00}(\mathbf{U}, |\hat{A}(\mathbf{U}, \alpha, J)|, \alpha) \cap L_0(\mathbf{U}, |\hat{A}(\mathbf{U}, \alpha, J)|, \alpha, J)$. Therefore:

$$\begin{aligned}
\mathbb{E}\left(\frac{|\hat{A}(\mathbf{U}, \alpha, J) \cap A^c|}{|\hat{A}(\mathbf{U}, \alpha, J) \vee 1|}\right) &= \sum_{v \in V_0} \mathbb{E}\left(\frac{\mathbf{1}_{v \in L_{00}(\mathbf{U}, i_*, \alpha) \cap L_0(\mathbf{U}, i_*, \alpha, J)}}{|\hat{A}(\mathbf{U}, \alpha, J) \vee 1|}\right) \\
&\leq \sum_{v \in V_0} \mathbb{E}\left(\frac{\mathbf{1}_{v \in L_{00}(\mathbf{U}, |\hat{A}(\mathbf{U}, \alpha, J)|, \alpha) \cap L_0(\mathbf{U}, |\hat{A}(\mathbf{U}, \alpha, J)|, \alpha, J)}}{|\hat{A}(\mathbf{U}, \alpha, J) \vee 1|}\right) \\
&\leq \sum_{v \in V_{00}} \mathbb{E}\left(\frac{\mathbf{1}_{v \in L_{00}(\mathbf{U}, |\hat{A}(\mathbf{U}, \alpha, J)|, \alpha)}}{|\hat{A}(\mathbf{U}, \alpha, J) \vee 1|}\right) + \sum_{v \in V_0 \setminus V_{00}} \mathbb{E}\left(\frac{\mathbf{1}_{v \in L_0(\mathbf{U}, |\hat{A}(\mathbf{U}, \alpha, J)|, \alpha, J)}}{|\hat{A}(\mathbf{U}, \alpha, J) \vee 1|}\right).
\end{aligned}$$

It will be then sufficient to prove that for every $v \in V_{00}$:

$$\mathbb{E}\left(\frac{\mathbf{1}_{v \in L_{00}(\mathbf{U}, |\hat{A}(\mathbf{U}, \alpha, J)|, \alpha)}}{|\hat{A}(\mathbf{U}, \alpha, J) \vee 1|}\right) \leq \frac{\alpha}{|V|},$$

and for every $v \in V_0 \setminus V_{00}$:

$$\mathbb{E}\left(\frac{\mathbf{1}_{v \in L_0(\mathbf{U}, |\hat{A}(\mathbf{U}, \alpha, J)|, \alpha, J)}}{|\hat{A}(\mathbf{U}, \alpha, J) \vee 1|}\right) \leq \frac{\alpha}{J}.$$

For any $v \in V_{00}$:

$$\begin{aligned}\mathbb{E}\left(\frac{\mathbf{1}_{v\in L_{00}(\mathbf{U},|\hat{A}(\mathbf{U},\alpha,J)|,\alpha)}}{|\hat{A}(\mathbf{U},\alpha,J)\vee 1|}\right) &= \sum_{k=1}^{|V|}\mathbb{E}\left(\frac{\mathbf{1}_{v\in L_{00}(\mathbf{U},k,\alpha)}}{k}\mathbf{1}_{|\hat{A}(\mathbf{U},\alpha,J)|=k}\right)\\ &= \sum_{k=1}^{|V|}\frac{1}{k}\mathbb{P}\left(|\hat{A}(\mathbf{U},\alpha,J)|=k|T_v^{00}(\mathbf{U})\leq\frac{k\alpha}{|V|}\right)\mathbb{P}\left(T_v^{00}(\mathbf{U})\leq\frac{k\alpha}{|V|}\right)\\ &\leq \frac{\alpha}{|V|}\sum_{k=1}^{|V|}\mathbb{P}\left(|\hat{A}(\mathbf{U},\alpha,J)|=k|T_v^{00}(\mathbf{U})\leq\frac{k\alpha}{|V|}\right)\end{aligned}$$

At this point, one can rewrite the last sum as following:

$$\frac{\alpha}{|V|}\sum_{k=0}^{|V|-1}\mathbb{P}\left(|\hat{A}(\mathbf{U},\alpha,J)|\leq k+1|T_v^{00}(\mathbf{U})\leq\frac{k\alpha}{|V|}\right)-\mathbb{P}\left(|\hat{A}(\mathbf{U},\alpha,J)|\leq k|T_v^{00}(\mathbf{U})\leq\frac{k\alpha}{|V|}\right).$$

But notice that for every $k\geq 0$, we can rewrite the event $|\hat{A}(\mathbf{U},\alpha,J)|\leq k+1$ as $(T_v^{00}(\mathbf{U}),T_v^0(\mathbf{U}))_{v\in V}\in D_{k+1}$, where D_{k+1} is some nondecreasing set of $\mathbb{R}^{2|V|}$. Therefore, we can apply lemma 4.1.1, and state that the last sum is less than:

$$\frac{\alpha}{|V|}\sum_{k=0}^{|V|-1}\mathbb{P}\left(|\hat{A}(\mathbf{U},\alpha,J)|\leq k+1|T_v^{00}(\mathbf{U})\leq\frac{(k+1)\alpha}{|V|}\right)-\mathbb{P}\left(|\hat{A}(\mathbf{U},\alpha,J)|\leq k|T_v^{00}(\mathbf{U})\leq\frac{k\alpha}{|V|}\right).$$

And the last sum is immediately less than 1. For the term $\mathbb{E}\left(\frac{\mathbf{1}_{v\in L_0(\mathbf{U},|\hat{A}(\mathbf{U},\alpha,J)|,\alpha,J)}}{|\hat{A}(\mathbf{U},\alpha,J)\vee 1|}\right)$, the steps are identical and it suffices to notice that $\mathbb{P}\left(T_v^0(\mathbf{U})\leq\frac{k\alpha}{|V|}\right)\leq\frac{\alpha k}{J}$ for every

k, which finishes the proof.

□

4.1.2 Procedure with an estimated upper bound of J

Algorithm 5 requires the knowledge of (an upper bound) of the number J. As announced, we will know relax assumption $[A1]$, and assume instead that we have a procedure giving an estimator $\hat{J}(\mathbf{U}, \epsilon)$ of J satisfying $\mathbb{P}\left(\hat{J}(\mathbf{U}, \epsilon) \leq J\right) \leq \epsilon$. However, the relaxation comes with the cost of strengthening the PRDS assumption. More precisely, the assumptions that we make are the following:

- ⊙ A1. We assume that we have an estimator $\hat{J}(\mathbf{U}, \epsilon)$ of J satisfying $\mathbb{P}\left(\hat{J}(\mathbf{U}, \epsilon) \leq J\right) \leq \epsilon$. For a procedure providing such a J, we refer to section 2.3.

- ⊙ A2. We also assume that the random vector $\left((T_v^0(\mathbf{U}), T_v^{00}(\mathbf{U}))_{v \in V}, -R(\mathbf{U}, \alpha, \hat{J}(\mathbf{U}, \epsilon))\right)$ is PRDS (positive regression dependence on each of a subset) on the vector $(T_v^0(\mathbf{U}), T_v^{00}(\mathbf{U}))_{v \in V}$, where:

$$R(\mathbf{U}, \alpha, \hat{J}(\mathbf{U}, \epsilon)) = L_{00}(\mathbf{U}, \hat{J}(\mathbf{U}, \epsilon), \alpha - \epsilon) \cap L_0(\mathbf{U}, \hat{J}(\mathbf{U}, \epsilon), \alpha - \epsilon)$$

With this new assumption, we present the following procedure:

Algorithm 6 Benjamini-Hochberg coarse-to-fine procedure with an unknown J

1: Compute $\hat{J}(\mathbf{U}, \epsilon)$ such that $\mathbb{P}\left(\hat{J}(\mathbf{U}, \epsilon) \leq J\right) \leq \epsilon$
2: Compute $i_* = \max\{i \leq |V| : i \leq |L_{00}(\mathbf{U}, i, \alpha - \epsilon) \cap L_0(\mathbf{U}, i, \alpha - \epsilon, \hat{J}(\mathbf{U}, \epsilon))|\}$
3: Define the rejection set as:

$$\hat{A}(\mathbf{U}, \alpha, \hat{J}(\mathbf{U}, \epsilon)) = L_{00}(\mathbf{U}, i_*, \alpha - \epsilon) \cap L_0(\mathbf{U}, i_*, \alpha - \epsilon, \hat{J}(\mathbf{U}, \epsilon)).$$

Algorithm 6 allows to control the coarse to fine procedure at a desired level α. More precisely:

Theorem 4.1.2. *Under the assumptions [A1.] and [A2.], we have:*

$$\mathbb{E}\left(\frac{|\hat{A}(\mathbf{U}, \alpha, \hat{J}(\mathbf{U}, \epsilon)) \cap A^c|}{|\hat{A}(\mathbf{U}, \alpha, \hat{J}(\mathbf{U}, \epsilon)) \vee 1|}\right) \leq \alpha.$$

Proof. The key observation here is that $L_0(\mathbf{U}, i, \alpha - \epsilon, J_1) \subset L_0(\mathbf{U}, i, \alpha - \epsilon, J_2)$ if $J_1 \geq J_2$. Indeed:

$$\mathbb{E}\left(\frac{|\hat{A}(\mathbf{U}, \alpha, \hat{J}(\mathbf{U}, \epsilon)) \cap A^c|}{|\hat{A}(\mathbf{U}, \alpha, \hat{J}(\mathbf{U}, \epsilon)) \vee 1|}\right) \leq \mathbb{E}\left(\frac{|\hat{A}(\mathbf{U}, \alpha, \hat{J}(\mathbf{U}, \epsilon)) \cap A^c|}{|\hat{A}(\mathbf{U}, \alpha, \hat{J}(\mathbf{U}, \epsilon)) \vee 1|}\mathbf{1}_{\hat{J}(\mathbf{U}, \epsilon) > J}\right) + \mathbb{P}\left(\hat{J}(\mathbf{U}, \epsilon) \leq J\right),$$

which is less than:

$$\sum_{v \in V_0} \sum_{k=1}^{|V|} \mathbb{E}\left(\frac{\mathbf{1}_{v \in L_{00}(\mathbf{U}, k, \alpha - \epsilon) \cap L_0(\mathbf{U}, k, \alpha - \epsilon, \hat{J}(\mathbf{U}, \epsilon))}}{k} \mathbf{1}_{\hat{J}(\mathbf{U}, \epsilon) > J} \mathbf{1}_{|\hat{A}(\mathbf{U}, \alpha, \hat{J}(\mathbf{U}, \epsilon))| = k}\right) + \epsilon,$$

which in turn is less than:

$$\sum_{v\in V_0}\sum_{k=1}^{|V|}\mathbb{E}\left(\frac{\mathbf{1}_{v\in L_{00}(\mathbf{U},k,\alpha-\epsilon)\cap L_0(\mathbf{U},k,\alpha-\epsilon,J)}}{k}\mathbf{1}_{|\hat{A}(\mathbf{U},\alpha,\hat{J}(\mathbf{U},\epsilon))|=k}\right)+\epsilon\leq\alpha-\epsilon+\epsilon,$$

which ends the proof (the steps required to get the last inequality form the third line are identical to the known J case). $|\hat{A}(\mathbf{U},\alpha,\hat{J}(\mathbf{U},\epsilon))|\leq k$ is non decreasing because of assumption [A2.]. Unfortunately, it is not possible to prove this property without strengthening the PRDS assumption via [A2.] and one can argue that it is impossible in general to check [A2.].

□

4.1.3 Procedure with general dependency structure

In reality, it is possible to avoid making assumptions on the dependency structure of the statistics $(T_v^0(\mathbf{U}),T_v^{00}(\mathbf{U}))_{v\in V}$ and $R(\mathbf{U},\alpha,\hat{J}(\mathbf{U},\epsilon))$. In this part, we will present a family of coarse-to-fine procedures that will control the FDR under any dependency structure. More precisely, we only assume:

⊙ A. We assume that we have an estimator $\hat{J}(\mathbf{U},\epsilon)$ of J satisfying $\mathbb{P}\left(\hat{J}(\mathbf{U},\epsilon)\leq J\right)\leq\epsilon$.

Of course, these procedures will have the disadvantage of being more conservative

than the coarse-to-fine procedure presented above.[10] In this part we will let ν_{00} and ν_0 be any probability measures on the set $\{1, 2.., |V|\}$. Let us define, for every probability measures ν_{00} and ν_0, the functions $\beta_{00,\nu_{00}}$ and β_{0,ν_0} from $\{1, 2, ..., |V|\}$ to $\mathbb{R}_+$ as :

$$\beta_{00,\nu}(i) = \sum_{k=1}^{i} k\nu_{00}(k),$$

and

$$\beta_{0,\nu}(i) = \sum_{k=1}^{i} k\nu_0(k).$$

With these notations, we define for any ν_{00} and ν_0 the family of detection sets:

$$L_{00}(\mathbf{U}, i, \alpha, \nu) := \{v \in V : T_v^{00}(\mathbf{U}) \leq \frac{\alpha\beta_{00,\nu_{00}}(i)}{|V|}\},$$

and

$$L_0(\mathbf{U}, i, \alpha, J) := \{v \in V : T_v^{0}(\mathbf{U}) \leq \frac{\alpha\beta_{0,\nu_0}(i)}{J}\}.$$

Using these detection sets, the family of procedures we propose, which we refer to as coarse-to-fine ν-procedures, is defined as the following :

Algorithm 7 Coarse-to-fine ν-procedures.

1: Compute $\hat{J}(\mathbf{U}, \epsilon)$ such that $\mathbb{P}\left(\hat{J}(\mathbf{U}, \epsilon) \leq J\right) \leq \epsilon$
2: Compute $i_* = \max\{i \leq |V| : i \leq |L_{00}(\mathbf{U}, i, \frac{\alpha-\epsilon}{2}, \nu) \cap L_0(\mathbf{U}, i, \frac{\alpha-\epsilon}{2}, \hat{J}(\mathbf{U}, \epsilon))|\}$
3: Define the rejection set as:

$$\hat{A}(\mathbf{U}, \alpha, \hat{J}(\mathbf{U}, \epsilon), \nu) = L_{00}(\mathbf{U}, i_*, \frac{\alpha - \epsilon}{2}) \cap L_0(\mathbf{U}, i_*, \frac{\alpha - \epsilon}{2}, \hat{J}(\mathbf{U}, \epsilon)).$$

Assuming [A.], algorithm 7 controls the FDR at the desired level. Namely:

Theorem 4.1.3.

$$\mathbb{E}\left(\frac{|\hat{A}(\mathbf{U}, \alpha, \hat{J}(\mathbf{U}, \epsilon), \nu) \cap A^c|}{|\hat{A}(\mathbf{U}, \alpha, \hat{J}(\mathbf{U}, \epsilon)) \vee 1|}\right) \leq \alpha.$$

To, prove theorem 4.1.3, we will need the following central lemma:

Lemma 4.1.2 (See[10])**.** *Let c be any positive constant, ν any probability measure on the positive real line, Z a random variable that dominates a uniform on $[0, 1]$ and Z any random variable. Finally let β_ν be the function defined as:*

$$\beta_\nu(x) = \int_0^x y \; d\nu(y).$$

Therefore, we have:

$$\mathbb{E}\left(\frac{\mathbf{1}_{Z \leq c\beta_\nu(W)}}{W \vee 1}\right) \leq c$$

Proof of lemma 4.1.2. The key and simple idea is to write $\frac{1}{W \vee 1}$ as $\int_0^\infty \frac{1}{z^2} \mathbf{1}_{\{z \geq W \vee 1\}} \; dz$, so that:

$$\begin{aligned}
\mathbb{E}\left(\frac{\mathbf{1}_{Z\leq c\beta_\nu(W)}}{W\vee 1}\right) &= \mathbb{E}\left(\int_0^\infty \frac{1}{z^2}\mathbf{1}_{\{z\geq W\vee 1\}}\mathbf{1}_{Z\leq c\beta_\nu(W)}\ dz\right)\\
&\leq \mathbb{E}\left(\int_0^\infty \frac{1}{z^2}\mathbf{1}_{Z\leq c\beta_\nu(z)}\ dz\right)\\
&= \int_0^\infty \mathbb{P}\left(Z\leq c\beta_\nu(z)\right)\frac{1}{z^2}dz\\
&\leq c\int_0^\infty y\int_0^\infty \frac{1}{z^2}\mathbf{1}_{z\geq y}dzd\nu(y)\\
&= c
\end{aligned}$$

□

We will now prove theorem 4.1.3.

Proof.

$$\begin{aligned}
\mathbb{E}\left(\frac{|\hat{A}(\mathbf{U},\alpha,\hat{J}(\mathbf{U},\epsilon))\cap A^c|}{|\hat{A}(\mathbf{U},\alpha,\hat{J}(\mathbf{U},\epsilon))\vee 1|}\right) &\leq \mathbb{E}\left(\frac{|\hat{A}(\mathbf{U},\alpha,\hat{J}(\mathbf{U},\epsilon))\cap A^c|}{|\hat{A}(\mathbf{U},\alpha,\hat{J}(\mathbf{U},\epsilon))\vee 1|}\mathbf{1}_{\hat{J}(\mathbf{U},\epsilon)>J}\right)+\mathbb{P}\left(\hat{J}(\mathbf{U},\epsilon)\leq J\right)\\
&= \sum_{v\in V_{00}}\mathbb{E}\left(\frac{\mathbf{1}_{v\in L_{00}(\mathbf{U},i_*,\frac{\alpha-\epsilon}{2})}}{|\hat{A}(\mathbf{U},\alpha,\hat{J}(\mathbf{U},\epsilon))\vee 1|}\right)+\sum_{v\in V_0\setminus V_{00}}\mathbb{E}\left(\frac{\mathbf{1}_{v\in L_0(\mathbf{U},i_*,\frac{\alpha-\epsilon}{2},\hat{J}(\mathbf{U},\epsilon))}\mathbf{1}_{\hat{J}(\mathbf{U},\epsilon)>J}}{|\hat{A}(\mathbf{U},\alpha,\hat{J}(\mathbf{U},\epsilon))\vee 1|}\right)\\
&\quad+\epsilon
\end{aligned}$$

But for every $v \in V_{00}$:

$$\begin{aligned}\mathbb{E}\left(\frac{\mathbf{1}_{v\in L_{00}(\mathbf{U},i_*,\frac{\alpha-\epsilon}{2},\nu)}}{|\hat{A}(\mathbf{U},\alpha,\hat{J}(\mathbf{U},\epsilon),\nu)\vee 1|}\right) &\leq \mathbb{E}\left(\frac{\mathbf{1}_{v\in L_{00}(\mathbf{U},|\hat{A}(\mathbf{U},\alpha,\hat{J}(\mathbf{U},\epsilon),\nu)\vee 1|,\frac{\alpha-\epsilon}{2})}}{|\hat{A}(\mathbf{U},\alpha,\hat{J}(\mathbf{U},\epsilon),\nu)\vee 1|}\right)\\ &= \mathbb{E}\left(\frac{\mathbf{1}_{T_v^{00}(\mathbf{U})\leq\frac{(\alpha-\epsilon)\beta_{00,\nu}(|\hat{A}(\mathbf{U},\alpha,\hat{J}(\mathbf{U},\epsilon),\nu)|)}{2|V|}}}{|\hat{A}(\mathbf{U},\alpha,\hat{J}(\mathbf{U},\epsilon),\nu)\vee 1|}\right)\\ &\leq \frac{\alpha-\epsilon}{2|V|},\end{aligned}$$

where we used lemma 4.1.2 for the last inequality.

Now, let us take some $v \in V_0 \setminus V_{00}$. We first claim that if $J_2 \geq J_1$, then $L_0(\mathbf{U}, |\hat{A}(\mathbf{U},\alpha,\hat{J}(\mathbf{U},\epsilon),\nu)|, \alpha, J_2) \subset L_0(\mathbf{U}, |\hat{A}(\mathbf{U},\alpha,\hat{J}(\mathbf{U},\epsilon),\nu)|, \alpha, J_1)$. Indeed, for any $J_2 \leq J_1$:

$$\frac{\alpha\beta_{0,\nu_0}(|\hat{A}(\mathbf{U},\alpha,\hat{J}(\mathbf{U},\epsilon),\nu)|)}{J_2} \leq \frac{\alpha\beta_{0,\nu_0}(|\hat{A}(\mathbf{U},\alpha,\hat{J}(\mathbf{U},\epsilon),\nu)|)}{J_2}$$

This implies that :

$$\begin{aligned}\sum_{v\in V_0\setminus V_{00}}\mathbb{E}\left(\frac{\mathbf{1}_{v\in L_0(\mathbf{U},i_*,\frac{\alpha-\epsilon}{2},\hat{J}(\mathbf{U},\epsilon))}\mathbf{1}_{\hat{J}(\mathbf{U},\epsilon)>J}}{|\hat{A}(\mathbf{U},\alpha,\hat{J}(\mathbf{U},\epsilon))\vee 1|}\right) &\leq \sum_{v\in V_0\setminus V_{00}}\mathbb{E}\left(\frac{\mathbf{1}_{v\in L_0(\mathbf{U},|\hat{A}(\mathbf{U},\alpha,\hat{J}(\mathbf{U},\epsilon))|,\frac{\alpha-\epsilon}{2},J)}}{|\hat{A}(\mathbf{U},\alpha,\hat{J}(\mathbf{U},\epsilon))\vee 1|}\right)\\ &\leq \sum_{v\in V_0\setminus V_{00}}\mathbb{E}\left(\frac{\mathbf{1}_{T_v^0(\mathbf{U})\leq\frac{(\alpha-\epsilon)\beta_{0,\nu_0}(i)}{2J}}}{|\hat{A}(\mathbf{U},\alpha,\hat{J}(\mathbf{U},\epsilon))\vee 1|}\right)\\ &\leq \frac{(\alpha-\epsilon)}{2J},\end{aligned}$$

where here again we used lemma 4.1.2 for the last inequality, which terminates the proof.

□

4.2 Finite resampling procedure

In fact, we will skip details in this part since all the procedures developed in section are immediately applicable to the finite resampling case and it suffices generate i.i.d elements $\xi_1, \xi_2, ...\xi_K$ for any integer K, and replace for every $v \in V$, the statistics $T_v^0(\mathbf{U})$ and $T_v^{00}(\mathbf{U})$ by:

$$\hat{T}_v^0(\mathbf{U}) := \frac{1}{K}\sum_{k=1}^{K} \mathbf{1}_{\rho_v(\mathbf{U}) \leq \rho_v(\xi_k \odot \mathbf{U})},$$

and

$$\hat{T}_v^{00}(\mathbf{U}) := \frac{1}{K}\sum_{k=1}^{K} \mathbf{1}_{\min(\phi_V(\rho_v(\mathbf{U})), \phi_G(\rho_{g(v)}(\mathbf{U}))) \leq \min(\phi_V(\rho_v(\xi_k \odot \mathbf{U})), \phi_G(\rho_{g(v)}(\xi_k \odot \mathbf{U})))}.$$

With these natural modifications, we have the following theorems, which proofs are immediate:

Theorem 4.2.1. *Assume without loss of generality that* $\frac{K\alpha}{|V|}$ *and* $\frac{K\alpha}{J}$ *are integers. Then:*

- ⊙ *With the assumption of theorem 4.1.1, the finite resampling version of algorithm 5 controls the FDR at level* $\alpha + \frac{|V|+J}{K+1}$

- ⊙ *With the assumption of theorem 4.1.2, the finite resampling version of algorithm 6 controls the FDR at level* $\alpha + \frac{|V|+J}{K+1}$

- ⊙ *With the assumption of theorem 4.1.3, the finite resampling version of algorithm 7 controls the FDR at level* $\alpha + \frac{|V|+J}{K+1}$

4.3 Simulations

Similarly to the previous simulations , we used PLINK generated datasets, and the setting is identical to Section 2.5. We generated datasets with parameters from Section 2.5, iterating over $\alpha = 0.5$, 1, 5, 10, 20, 30, 40 and 50 in the Chinese restaurant process and considering four cases: (I) $\nu_G = 10$, no LD; (II) $\nu_G = 25$, no LD; (III) $\nu_G = 10$, LD = 0.8; (III) $\nu_G = 25$, LD = 0.8. In each case, we took the average over 50 simulations with $K = 10^6$. The only difference here with the setting of previous chapters is that the active set has size 50 instead of 25. In total, We compare 11 different methods. The first 10 methods consist in looking at 5 one level methods and their equivalent coarse-to-fine version, and the 11th method is the Bonferroni-Holm estimator. The first method considered is the Benjamini-Hochberg method and its coarse-to-fine version with an estimated $\hat{J}$. Then we looked at the family of procedures controlling the FDR for all dependence setting proposed in[10] and described in 1.3.

Table 4.1: Cell size 10. Independent SNPs. Average number of detections for single level procedures.

Benjamini Hochberg	ν-procedure $\beta = 0.1$	ν-procedure $\beta = 0.5$	ν-procedure $\beta = 2$	ν-procedure Uniform	Bonferroni
25.31	15.86	12.05	11.94	12.16	12.14

We took $\nu(k) \propto \exp(-\beta k)$ for $\beta = 0.1, 0.5$ and 2. We then looked at their coarse-to-fine versions with $\nu_0 = \nu_{00} = \nu$ in each of the three cases. Finally, we took ν a uniform on $\{1, 2, ..., |V|\}$, $\nu_{00} = \nu$ and ν_0 a uniform on $\{1, 2, ..., \hat{J}\}$. At this stage, one can argue that the support of ν_0 depends on $\hat{J}$ and there is no guarantee that the FDR will be controlled for any setting. In reality, this is not the case and one just needs to notice that $\frac{\beta_{0,\nu_J}(k)}{J}$ is decreasing in J for every integer k where ν_J is the uniform distribution on $\{1, 2, ..., J\}$. Using this remark, proving that the FDR is always controlled is identical to the proof of 4.1.3. In all of the simulations, the desired FDR level to control is 0.1. Even if there is no theoretical guarantee for the coarse-to-fine Benjamini-Hochberg method, we mention that in practice, the FDR is controlled at the desired level. Finally, let us mention that when the clustering assumption is true, the ν-coarse to fine procedure with β equals to 0.1 has a performance that is comparable to the Benjamini-Hochberg method. Recall that this procedure presents theoretical guarantees that the FDR is controlled under any dependence setting.

Table 4.2: Cell size 10. Independent SNPs. Average number of detections for coarse-to-fine procedures

Value of α	CTF Benjamini Hochberg	CTF ν-procedure $\beta = 0.1$	CTF ν-procedure $\beta = 0.5$	CTF ν-procedure $\beta = 2$	CTF ν-procedure Uniform
0.5	34.26	27.18	21.74	18.62	12.68
1	32.78	25.34	19.8	16	12.18
5	29.88	22.16	8.9	15.28	11.24
10	26.49	19.04	15.76	13.26	10.2
20	23.42	15.68	13.56	11.52	8.78
30	20.09	13.54	12.32	10.8	8
40	16.48	11.51	11.26	9.44	7.22
50	14.26	10.96	10.54	9.98	6.76

Table 4.3: Cell size 10. Correlated SNPs. Average number of detections for single level procedures

Benjamini Hochberg	ν-procedure $\beta = 0.1$	ν-procedure $\beta = 0.5$	ν-procedure $\beta = 2$	ν-procedure Uniform	Bonferroni
23.74	15.65	12.15	12.07	12.46	12.45

Table 4.4: Cell size 10. Correlated SNPs. Average number of detections for coarse-to-fine procedures

Value of α	CTF Benjamini Hochberg	CTF ν-procedure $\beta = 0.1$	CTF ν-procedure $\beta = 0.5$	CTF ν-procedure $\beta = 2$	CTF ν-procedure Uniform
0.5	29.62	24.56	21.6	16.62	11.82
1	29.04	24.12	20.78	16.94	11.7
5	26.64	22.06	18.82	15.56	11.22
10	24.44	20.32	17.06	14.68	10.94
20	22.06	18.08	15.02	13.2	10.16
30	20	16.3	13.44	11.46	9.6
40	18.14	11.86	12.16	10.1	9.12
50	17.97	11.66	12.16	8.22	8.48

Table 4.5: Cell size 25. Independent SNPs. Average number of detections for single level procedures

Benjamini Hochberg	ν-procedure $\beta = 0.1$	ν-procedure $\beta = 0.5$	ν-procedure $\beta = 2$	ν-procedure Uniform	Bonferroni
25.48	16.01	12.37	12.08	12.36	12.37

Table 4.6: Cell size 25. Independent SNPs. Average number of detections for coarse-to-fine procedures

Value of α	CTF Benjamini Hochberg	CTF ν-procedure $\beta = 0.1$	CTF ν-procedure $\beta = 0.5$	CTF ν-procedure $\beta = 2$	CTF ν-procedure Uniform
0.5	35.12	29.66	21.74	19.36	12.62
1	33.08	27.54	20.22	17.42	11.64
5	28.34	22.56	17.18	14.36	10.28
10	24.42	18.66	14.86	12.3	8.9
20	19.92	14.54	12.2	10	7.36
30	16.46	11.26	10.48	8.68	6.34
40	13.26	8.7	9.06	7.84	5.16
50	10.7	6.82	7.66	7.18	5.48

Table 4.7: Cell size 25. Correlated SNPs. Average number of detections for single level procedures

Benjamini Hochberg	ν-procedure $\beta = 0.1$	ν-procedure $\beta = 0.5$	ν-procedure $\beta = 2$	ν-procedure Uniform	Bonferroni
23.64	15.36	11.95	12	12.13	11.7

Table 4.8: Cell size 25. Correlated SNPs. Average number of detections for coarse-to-fine procedures

Value of α	CTF Benjamini Hochberg	CTF ν-procedure $\beta = 0.1$	CTF ν-procedure $\beta = 0.5$	CTF ν-procedure $\beta = 2$	CTF ν-procedure Uniform
0.5	31.86	26.44	23.24	20.04	14.8
1	31	25.5	21.88	18.74	12.76
5	26.58	21.76	19.12	16.32	10.94
10	22.56	18.2	16.34	13.84	9.12
20	18.48	14.42	13.4	11.02	7.14
30	15.02	11.24	10.72	8.88	6.32
40	11.74	7.98	7.88	6.36	5.38
50	11.5	6.96	6.11	5.2	5.06

Chapter 5

Conclusion and discussion

Given a partition of the space of hypotheses, the basic assumption which allows the coarse-to-fine multiple testing algorithm to obtain greater power than single level approaches at the same FWER or FDR level is that the distribution of the numbers of active hypotheses across the cells of the partition is non-uniform. One can separate our work into three main parts. In the first part, we gave a coarse-to-fine version of the Bonferroni procedure based on a resampling method . At the beginning, we assumed that we have access to asymptotic resampling and to the number of non-active indices J in active cells and showed that we control the FWER. We then modified this algorithm in order to control the FWER in the finite resampling case. Later on, we relaxed the known J assumption, replacing it by an estimated upper bound $\hat{J}$. We also derived a coarse-to-fine Bonferroni method for a Gaussian linear parametric model. The main interest of this model was the possibility to have estimates of

the power of detection as a function of the thresholds under known alternatives and compare, through simulations, the power of detection of the coarse-to-fine method using the default parameters to the method using the optimal parameters. In the second part, we argued that the Bonferroni coarse-to-fine method wasn't well suited when there is high dependence between the tests. Using the so called zero assumption that we introduced to compute $\hat{J}$, we replaced it by a dependence adapted coarse-to-fine method that proved to be superior to the previous coarse-to-fine method when we look at the simulations. We first described the procedure in the infinite resampling case, before modifying the procedure in three different ways suited for the finite resampling framework. In the third part of the manuscript, we switch gears and focus on coarse-to-fine methods controlling the FDR. We began proposing a coarse-to-fine version of the Benjamini-Hochberg procedure under a known J and PRDS assumptions. We saw that it is possible to replace J by an estimated $\hat{J}$, provided that we make an assumption slightly stronger than the PRDS property. Finally, we derived a family of coarse-to-fine algorithms controlling the FDR under any given dependence between the test, that is however more conservative than the coarse-to-fine Benjamini-Hochberg procedure. Let us mention that for all the procedures in this third part, the modification required between the infinite resampling case and the finite resampling case was immediate.

All of the coarse-to-fine procedures mentioned in these three parts showed to be efficient under the clustering assumption. This scenario was motivated by the

situation encountered in genome-wide association studies, where the hypotheses are associated with genetic variations (e.g., SNPs), each having a location along the genome, and the cells are associated with genes. In principle, our coarse-to-fine procedure will then detect more active variants to the extent that these variants cluster in genes. Of course, this extent will depend in practice on many factors, including effect sizes, the representation of the genotype (i.e., the choice of variants to explore) as well as the phenotype, and complex interactions within the genotype. It may be very difficult and uncommon to know anything specific about the expected nature of the combinatorics between genes and variants. Looking ahead, we have only analyzed the coarse-to-fine approach for the simplest case of two-levels and a true partition, i.e., non-overlapping cells. The methods for controlling the FWER for both the parametric and non-parametric cases generalize naturally to multiple levels assuming nested partitions. However, one will need to be careful when taking into account the finite resampling case, checking that the number K required does not grow with the number of levels. In reality, except for the coarse-to-fine Bonferroni method, we claim that the increase of K as a function of the number of levels is negligible, and the generalization in these cases will be easy. The other analytical challenge is to generalize the coarse-to-fine approach to overlapping cells, even for two levels: while our methods for controlling the FDR and FWER remain valid, they are likely to become overly conservative if cells overlap (however, one could artificially create the partitions by imposing the constraint of assigning at most one

cell to an index). This case is of particular interest in applications, where genes are grouped into overlapping "pathways." For example, in Systems Biology, cellular phenotypes, especially complex diseases such as cancer, are studied in the context of these pathways and mutated genes and other abnormalities are in fact known to cluster in pathways; indeed, this is the justification for a pathway-based analysis. Hence the clustering properties may be stronger for variants or genes in pathways than for variants in genes.

Bibliography

[1] W. S. Bush and J. H. Moore, "Genome-wide association studies," *PLoS computational biology*, vol. 8, no. 12, p. e1002822, 2012.

[2] J. W. Tukey, "Comparing individual means in the analysis of variance," *Biometrics*, pp. 99–114, 1949.

[3] K. R. Gabriel, "Simultaneous test procedures–some theory of multiple comparisons," *The Annals of Mathematical Statistics*, vol. 40, no. 1, pp. pp. 224–250, 1969. [Online]. Available: http://www.jstor.org/stable/2239213

[4] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979. [Online]. Available: http://www.jstor.org/stable/4615733

[5] Y. Hochberg, "A sharper Bonferroni procedure for multiple tests of significance," *Biometrika*, vol. 75, no. 4, p. 800, 1988. [Online]. Available: http://biomet.oxfordjournals.org/content/75/4/800.short

BIBLIOGRAPHY

[6] P. Good, *Permutation tests: a practical guide to resampling methods for testing hypotheses*. Springer Science & Business Media, 2013.

[7] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995. [Online]. Available: http://www.jstor.org/stable/2346101

[8] Y. Benjamini and D. Yekutieli, "The control of the false discovery rate in multiple testing under dependency," *Annals of Statistics*, pp. 1165–1188, 2001. [Online]. Available: http://www.jstor.org/stable/2674075

[9] Y. Benjamini, A. M. Krieger, and D. Yekutieli, "Adaptive linear step-up procedures that control the false discovery rate," *Biometrika*, vol. 93, no. 3, p. 491, 2006. [Online]. Available: http://biomet.oxfordjournals.org/content/93/3/491.short

[10] G. Blanchard, E. Roquain *et al.*, "Two simple sufficient conditions for fdr control," *Electronic journal of Statistics*, vol. 2, pp. 963–992, 2008.

[11] J. D. Storey, J. E. Taylor, and D. Siegmund, "Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 66, no. 1, pp. 187–205, 2004. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9868.2004.00439.x/full

BIBLIOGRAPHY

[12] B. Efron, *The jackknife, the bootstrap and other resampling plans.* SIAM, 1982.

[13] B. Efron and R. J. Tibshirani, *An introduction to the bootstrap.* CRC press, 1994.

[14] A. C. Davison and D. V. Hinkley, *Bootstrap methods and their application.* Cambridge university press, 1997, vol. 1.

[15] S. K. Sarkar, "Two-stage stepup procedures controlling fdr," *Journal of Statistical Planning and Inference*, vol. 138, no. 4, pp. 1072–1084, 2008.

[16] E. Candes, Y. Fan, L. Janson, and J. Lv, "Panning for gold: Model-free knockoffs for high-dimensional controlled variable selection," *arXiv preprint arXiv:1610.02351*, 2016.

[17] R. F. Barber and E. J. Candes, "A knockoff filter for high-dimensional selective inference," *arXiv preprint arXiv:1602.03574*, 2016.

[18] D. Brzyski, C. B. Peterson, P. Sobczyk, E. J. Candès, M. Bogdan, and C. Sabatti, "Controlling the rate of gwas false discoveries," *Genetics*, vol. 205, no. 1, pp. 61–75, 2017.

[19] E. L. Lehmann and J. P. Romano, "Generalizations of the familywise error rate," *Annals of Statistics*, pp. 1138–1154, 2005.

[20] T. Dickhaus, "Simultaneous statistical inference," in *With applications in the life sciences.* Springer, 2014.

BIBLIOGRAPHY

[21] B. Efron and R. Tibshirani, "Empirical bayes methods and false discovery rates for microarrays," *Genetic epidemiology*, vol. 23, no. 1, pp. 70–86, 2002.

[22] B. Efron, *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*. Cambridge University Press, 2012, vol. 1.

[23] A. Corvin, N. Craddock, and P. F. Sullivan, "Genome-wide association studies: a primer," *Psychological Medicine*, vol. 40, no. 07, pp. 1063–1077, 2010.

[24] B. L. Fridley, "Bayesian variable and model selection methods for genetic association studies," *Genetic Epidemiology*, vol. 33, no. 1, pp. 27–37, 2009.

[25] D. J. Balding, "A tutorial on statistical methods for population association studies," *Nature Reviews Genetics*, vol. 7, no. 10, pp. 781–791, 2006.

[26] J. N. Hirschhorn and M. J. Daly, "Genome-wide association studies for common diseases and complex traits," *Nature Reviews Genetics*, vol. 6, no. 2, pp. 95–108, 2005.

[27] O. L. Griffith, S. B. Montgomery, B. Bernier, B. Chu, K. Kasaian, S. Aerts, S. Mahony, M. C. Sleumer, M. Bilenky, M. Haeussler *et al.*, "Oreganno: an open-access community-driven resource for regulatory annotation," *Nucleic acids research*, vol. 36, no. suppl_1, pp. D107–D113, 2007.

[28] M. Li, C. Li, and W. Guan, "Evaluation of coverage variation of snp chips for

genome-wide association studies," *European Journal of Human Genetics*, vol. 16, no. 5, pp. 635–643, 2008.

[29] K. Kelley and K. J. Preacher, "On effect size." *Psychological methods*, vol. 17, no. 2, p. 137, 2012.

[30] H. Huang, P. Chanda, A. Alonso, J. S. Bader, and D. E. Arking, "Gene-based tests of association," *PLoS Genetics*, vol. 7, no. 7, p. e1002177, 2011.

[31] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig *et al.*, "Gene ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[32] G. Blanchard and D. Geman, "Hierarchical testing designs for pattern recognition," *Annals of Statistics*, pp. 1155–1202, 2005.

[33] N. Meinshausen, "Hierarchical testing of variable importance," *Biometrika*, vol. 95, no. 2, pp. 265–278, 2008.

[34] Y. Gavrilov, Y. Benjamini, and S. K. Sarkar, "An adaptive step-down procedure with proven fdr control under independence," *The Annals of Statistics*, pp. 619–629, 2009.

[35] T. Dickhaus, K. Straßburger, D. Schunk, C. Morcillo-Suarez, T. Illig, and A. Navarro, "How to analyze many contingency tables simultaneously in genetic

association studies," *Statistical applications in genetics and molecular biology*, vol. 11, no. 4, 2012.

[36] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. De Bakker, M. J. Daly *et al.*, "Plink: a tool set for whole-genome association and population-based linkage analyses," *The American Journal of Human Genetics*, vol. 81, no. 3, pp. 559–575, 2007.

[37] D. J. Aldous, "Exchangeability and related topics," in *École d'Été de Probabilités de Saint-Flour XIII, 1983*. Springer, 1985, pp. 1–198.

[38] V. N. Vapnik and A. Y. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," in *Measures of complexity*. Springer, 2015, pp. 11–30.

[39] L. Devroye, L. Györfi, and G. Lugosi, *A probabilistic theory of pattern recognition*. Springer Science & Business Media, 2013, vol. 31.

[40] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning*. MIT press, 2012.

Curriculum Vitae/Bio

⊙ Kamel Lahouel

⊙ 307B Clark Hall 3400 N Charles St.

⊙ Baltimore, MD 21200 Email: klahoue1@jhu.edu

Personal Information:

⊙ Date of birth: 03/29/1988.

⊙ Place: Tunis, Tunisia

Education:

⊙ 2013- present: Ph.D. student in Applied Mathematics and Statistics, Whiting School of Engineering, Johns Hopkins University, Baltimore, USA. Passed the Introductory Exam in January 2013 -

⊙ 2011-2012 : Ecole Normale Suprieure (ENS) de Cachan, Cachan, France, Third Year, Mathematics Department, M2 Vision Apprentissage (MVA), Master's degree, graduated with Honor

BIBLIOGRAPHY

- 2010-2011: Second Year at ENS Cachan. Degree obtained: M1 (Highest distinction: very good) - 2009-2010: First Year at ENS Cachan. Degree obtained: L3 (B.S. Distinction: good)
- 2008-2009 : MP* Lyce Condorcet, Paris. Admission to Ecole Normale Suprieure (ENS) Cachan, France
- 2007-2008 : MP* Lyce Condorcet, Paris - 2006-2007 : MPSI Lyce Condorcet, Paris
- 2005-2006: Lyce El Menzah IX, Tunis, Tunisia. Baccalaurat in Mathematics (Distinction: very good)

Awards:

- Full scholarship for PhD studies (four years) in Applied Mathematics and Statistics, Johns Hopkins University
- 2013: The Centennial Fellowship Prize for the Academic Year 2013-2014
- 2014: Counselman Fellowship
- 2015 : Counselman Fellowship

RESEARCH:

- 2015 to present: Tumorigenesis modeling
- 2015: Non-Adaptive Policies for 20 Questions Target Localization

BIBLIOGRAPHY

- 2012 to present: "Coarse to fine methods in multiple testing: application to detection of active variants in the human genome", under the direction of Professors D. Geman and L. Younes, Center of Imaging Science, Johns Hopkins University

- 2011: "A statistical test of translation between densities of two spatial Poisson Processes", Imagine, ENPC, under the direction of Pr. Arnak Dalalyan, ENS, Cachan, France

- 2010: "A Machine Learning challenge (SVRT)", Student joint research project, ENS Cachan, CMLA, under the direction of Pr. Alain Trouv

- 2009 : "A stochastic macroeconomic model: simulation of productivity shocks", TIPE Admission Exam to "Grandes Ecoles", Paris, France

Languages:

- Arabic: spoken and written.

- French: spoken and written.

- English: spoken and written

Computer Skills

- Microsoft Office: Word, Excel, Power Point - Latex

- Matlab -R -Python

BIBLIOGRAPHY

- Maple

EXTRA-CURRICULUM ACTIVITIES:

- Member of a social association helping to create jobs in disadvantaged regions in Tunisia (2010-2012)

- Sports practiced: soccer